Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática
Facultad de Ciencias
Departamento de Matemáticas

Análisis de Datos Funcionales, Clasificación y Selección de Variables

Trabajo Fin de Máster para la consecución de los titulos de
Máster en Ingeniería informática y Telecomunicaciones y
Máster en Matemáticas y Aplicaciones.

Por

José Luis Torrecilla Noguerales

bajo la dirección de

Antonio Cuevas González

Carlos Santa Cruz Fernández

Madrid, 27 de septiembre de 2010

Resumen

Los datos de alta dimensión y funcionales están ganando importancia en el campo de la clasificación debido a los avances técnicos que permiten su captura y tratamiento. Además, este tipo de datos son indicadores de elementos tan fundamentales como la salud o la economía. Todo esto hace que su estudio esté en auge.

Tradicionalmente, la aproximación a estos problemas se ha llevado a cabo desde dos puntos de vista: el más matemático, basado en modelos poblacionales subyacentes, y el más empírico de la minería de datos. Mientras la matemática, a través del Análisis de Datos Funcionales (FDA), estudia las funciones como objetos matemáticos, teniendo en cuenta sus propiedades y fundamentos teóricos, la minería de datos ve vectores de alta dimensión que trata mediante algoritmos propios atendiendo distintos criterios. Así, se ha considerado interesante abordar una amplia revisión bibliográfica del FDA hasta la actualidad y profundizar en el estudio de estos distintos enfoques viendo sus semejanzas y diferencias y vislumbrando posibles interacciones. En este último punto se pondrá especial atención a la utilización de métodos de selección de variables con datos funcionales, aspecto apenas explorado hasta la fecha. Todo ello en el marco de la clasificación supervisada.

Índice general

Capítulo 1

Introducción

La utilización de funciones de variable real (por ejemplo, el tiempo) es cada vez más habitual en diferentes campos de las ciencias experimentales y sociales, desde la medicina, con electrocardiogramas o electroencefalogramas, a la economía, donde recogen la evolución de índices o activos, pasando por otros muchos como la meteorología o el tratamiento de imágenes o señales.

Tradicionalmente, las propiedades de las funciones han sido objeto de estudio de los matemáticos dentro del análisis funcional con un enfoque eminentemente teórico. Sin embargo, actualmente, las mejoras tecnológicas permiten la captura y tratamiento de este tipo de datos haciendo interesante el estudio de la información que encierran. Como se ha dicho, las funciones modelan muchos procesos importantes, y extraer adecuadamente la información que contienen las curvas puede ser extremadamente útil. De este modo, se podría, por ejemplo, ver si una persona está sana, cómo van a evolucionar los mercados financieros o crear sistemas de reconocimiento de voz eficaces. Los métodos matemáticos orientados a estos fines se engloban bajo el epígrafe FDA (Análisis de Datos Funcionales) [81, 35, 79]. En FDA se consideran las propiedades de estos objetos (no son meros vectores de coordenadas) para desarrollar métodos eficaces y consistentes para hacer regresión, clasificación, predicción, proyección, etc. [3, 5, 10, 68, 22] Este trabajo se interesará especialmente la clasificación supervisada, estudiando distintas soluciones del problema en su versión funcional, por ejemplo, KNN, SVM, núcleos o medidas de profundidad [34, 87, 24]. Si bien, por el gran interés que despierta, lo novedoso del campo y la gran cantidad de aplicaciones que tiene, el núcleo de este trabajo será una extensa revisión bibliográfica del análisis de datos funcionales hasta la actualidad.

Por otro lado, la minería de datos tiene muchas similitudes con FDA. La minería de datos engloba todas aquellas técnicas encaminadas a la extracción de información no trivial que reside en un conjunto de datos. Para ello se apoya fuertemente en la inteligencia artificial y en la estadística,

creando modelos para predecir y clasificar, aunque desde otro punto de vista. Aquí, como veremos, un dato funcional es un mero vector y el énfasis se pone en los aspectos computacionales y algorítmicos, más que en el modelo probabilístico teórico que pueda subyacer en la obtención de los datos. Una buena referencia para profundizar en el área en el marco de la clasificación de patrones es [29] y con un enfoque más general [50, 12].

Hasta ahora, las funciones han sido estudiadas principalmente por matemáticos, muchas veces desde planteamientos puramente teóricos, mientras que la minería de datos, desarrollada especialmente por ingenieros, ha olvidado la naturaleza de estos objetos matemáticos complejos sin tener en cuenta (y aprovechar) sus propiedades. Uno de los objetivos de este trabajo es poner en común estos distintos conocimientos aplicando una técnica de la minería de datos hasta ahora no utilizada en FDA a datos funcionales.

La característica principal que diferencia el Análisis de Datos Funcionales del Multivariado propio de la minería de datos es la dimensión infinita de las funciones. Este aspecto distintivo es, a su vez, el principal escollo para tratar estos datos computacionalmente. Para abordar el problema se tiene que recurrir a una discretización de la función o a su expresión en una base reducida del espacio de funciones, con la consiguiente pérdida de información [81].

El problema de la alta dimensionalidad es, a la vez, crucial en el campo de la minería de datos, hasta el punto de hablarse de la maldición de la dimensionalidad (termino acuñado por Bellman en 1961), y en particular en la clasificación, constituyendo, la reducción de esta dimensionalidad, un punto de creciente interés por distintos motivos: aumenta la eficacia de los métodos, muestra estructuras ocultas, permite abordar mayores problemas o reduce el coste [12, 63]

Hasta ahora, una solución utilizada en ambas áreas es la reducción de dimensiones previa a la clasificación mediante proyecciones del espacio original (la discretización o las coordenadas en la nueva base) [35, 81]. Al hacer una reducción de dimensiones, se pueden tomar más puntos en la malla inicial y "perder" menos información, si bien los métodos de reducción por proyección como ICA (Análisis de Componentes Independientes), PCA (Análisis de Componentes Principales) o PLS (Mínimos Cuadrados Parciales) [65, 17, 62, 94], plantean algunos inconvenientes: estos algoritmos vuelven a ser costosos (diagonalización o inversión de matrices), puede cuestionarse el sentido de estas variables sintéticas para el problema, como en el PLS [35].

Otras técnicas de reducción de dimensionalidad son las que se engloban bajo el nombre de selección de variables. Estos métodos se basan en la selección de algunas de las variables originales del problema, de tal forma que sean las más representativas en función de ciertos criterios como la

correlación con la clase, la información mutua, o la redundancia entre variables. [42, 89, 63, 43]. La selección de variables, utilizada inicialmente en problemas de genética con pocas muestras y alta dimensionalidad, está extendiéndose en el campo de reconocimiento de patrones, siendo quizá su principal exponente el método de Mínima Redundancia y Máxima Relevancia (mRMR) [28].

La aplicación de estos métodos, más o menos generalizados en la minería de datos, al análisis funcional permitiría realizar discretizaciones más finas, eliminando a priori menos información sobre la curva, y tendría una interpretación muy clara: del mismo modo que nosotros miraríamos sólo la zona en la que las gráficas de las distintas clases difieren para clasificarlas, estos métodos encontrarían los instantes en los que las muestras de las clases son más separables. De este modo nos quedaríamos con un subconjunto mucho más pequeño y representativo.

Uno de los objetivos de este trabajo es relacionar las técnicas de selección de variables con las propiedades de los datos funcionales y en especial, aplicar métodos de selección de variables en conjuntos de datos funcionales y estudiar los resultados contrastándolos con algunos métodos de proyección generales, ya que la aplicación de selección de variables a problemas funcionales es un nuevo enfoque en la clasificación funcional que podría permitir discretizaciones muy finas emulando la dimensión infinita del problema original y en un tiempo razonable. Además, se podrían estudiar los comportamientos de algoritmos como mRMR como métodos de selección de variables graduables en función de distintos criterios y contrastar los resultados con otros como PLS.

Concluyendo, el Análisis de Datos Funcionales y la reducción de dimensiones son áreas con una gran potencialidad en cuanto a sus aplicaciones, constituyen campos en plena ebullición como muestran las fechas de publicación de muchos de los artículos referenciados, y la combinación de los planteamientos matemáticos con los propios de la ingeniería constituye un enfoque novedoso.

De este modo, el Trabajo Fin de Máster comienza con una visión general a nivel de estado del arte en el campo del Análisis de Datos Funcionales. Éste es un punto fundamental, y representa el grueso del trabajo. A través de las principales monografías sobre el tema [81, 35] y multitud de artículos se exponen los conceptos básicos y las dificultades que se presentan. Entre otros, la dimensión infinita, la alta correlación, los espacios de funciones, las herramientas analíticas en dichos espacios o representación finita de las funciones. Para después profundizar en los problemas: Regresión, Análisis de Componentes Principales y Clasificación. Descender a mínimos detalles se haría imposible debido a la magnitud de cada uno de estos aspectos, pero se verán las ideas fundamentales de todos ellos comentando su evolución. Además, se profundizará en la clasificación supervisada, estudiando varios de los algoritmos concretos, ya sean genuinos o adaptaciones del caso multivariado.

Habiendo estudiado el enfoque matemático y original para estos datos se estudiarán las soluciones tradicionales que ofrece la minería de datos (multivariadas) para problemas de alta dimensionalidad [94, 6, 43, 89]. Cualquiera que sea el enfoque utilizado, las funciones acaban manejándose en la práctica en una versión discretizada. La diferencia entre los enfoques mas "matemáticos" y los más "computacionales" (o de minería de datos) es que en los primeros, la motivación de los procedimientos es funcional y la discretización aparece como una simple aproximación. En este último área, las soluciones se dividen en dos bloques: reducción por proyección con métodos como PLS o PCA (ya utilizados en FDA), y la selección de variables. Ésta no involucra proyecciones ni componentes sintéticas y no ha sido aplicado aún en FDA. Para ello se estudiarán distintos métodos, con especial atención a mRMR y se realizará una breve discusión sobre la integración de los mismos.

Finalmente, el trabajo concluye con una serie de experimentos con distintos clasificadores, comparando los resultados obtenidos con los métodos hasta ahora utilizados en FDA con los obtenidos aplicando técnicas de selección de variables a los datos funcionales para valorar su posible utilización en el campo.

Capítulo 2

Análisis de Datos Funcionales (FDA)

En este capítulo se hace un resumen de la bibliografía estudiada con el objetivo de presentar el campo del Análisis de Datos funcional, llamado FDA por sus siglas en inglés. Ésta es una rama de las matemáticas, en concreto de la estadística, que estudia y analiza la información contenida en curvas, superficies, o cualquier elemento que varía sobre un continuo, generalmente el tiempo. Las particularidades de estos objetos, las funciones, hacen que surja todo un aparato de definiciones, teoremas y herramientas para poder estudiarlas correctamente. Por ello, el primer paso será definir y comentar los conceptos básicos, las características y peculiaridades de este tipo de datos. Una vez definido el marco y vistas las propiedades y dificultades a las que se enfrenta el FDA, haciendo especial hincapié en la representación, se resumen a grandes rasgos los principales problemas del área: regresión, análisis de componentes principales y clasificación. Estos tres puntos concentran la mayor parte de la investigación en el análisis de datos funcionales, siendo perfectamente representativos [68]. Para todos ellos se expondrá el problema y los objetivos, presentando algunas soluciones a nivel de estado del arte. Las limitaciones en extensión y alcance hacen que sólo profundicemos más en el caso de la clasificación supervisada, objetivo del trabajo, mostrando detalles de algunos algoritmos.

La explosión tecnológica ha hecho del FDA una disciplina emergente dentro de la estadística con multitud de publicaciones y estudios. Prueba de este interés es la creciente presencia en revistas del ramo y la creación de un grupo de trabajo específico dentro del ERCIM (European Research Consortium for Informatics and Mathematics). De entre toda la literatura disponible sobresalen como referencias básicas los libros de Ramsay y Silverman [81] y de Ferraty y Vieu [35] tratando muchos de los problemas básicos de la estadística funcional, en el segundo caso desde el punto de vista no-paramétrico. Ramsay y Silverman también tienen un segundo libro de carácter más aplicado en el que se estudian soluciones a problemas sobre conjuntos de datos concretos [80]. Y recientemente se ha publicado un libro de Ramsay, Hooker y Graves centrándose en los aspectos

computacionales en R y MATLAB del FDA [79]. Además se puede obtener información adicional en la página mantenida por Ramsay, http://www.functionaldata.org, y en la del grupo de Ferraty y Vieu, el grupo STAPH, http://www.lsp.ups-tls.fr/staph.

2.1. Conceptos Básicos

Antes de discutir las propiedades y los problemas del FDA es necesario fijar algunos conceptos de una manera más formal. Para ello se han estudiado las monografías referencia del área [81, 35], eligiendo como patrón en cuanto al lenguaje la de Ferraty y Vieu, y un libro clásico en análisis funcional [59]. Algunos de estos conceptos son muy básicos pero se ha considerado oportuno introducirlos por razones de completitud.

La primera cuestión es definir que se entiende por una función o variable funcional. En general, cualquier observación que varíe sobre un continuo se puede tomar por un dato funcional, desde un electrocardiograma a las temperaturas de una ciudad. En la práctica, estos sucesos son recogidos por máquinas que toman muestras de una determinada variable aleatoria en distintos instantes de tiempo dentro de un cierto rango (t_{min}, t_{max}). De este modo, una observación puede expresarse mediante la familia $\{X(t_j)\}_{j=1,\cdots,J}$. Actualmente, los avances en computación hacen posible hacer la rejilla cada vez más fina (uno de los motivos de la relevancia del FDA) de modo que podríamos considerar estas muestras como observaciones de la familia continua $\mathcal{X} = \{X(t); t \in (t_{min}, t_{max})\}$. Así, una *Variable Funcional* es aquella variable aleatoria $\mathcal{X}$ que toma valores en un espacio funcional (de dimensión infinita). Una observación χ de $\mathcal{X}$ se llama *Dato Funcional*.

Nos estamos centrando en el caso de una curva unidimensional, que es el más común, pero la definición es general, $\mathcal{X} = \{X(t); t \in T)\}$ y por tanto, $\chi = \{x(t); t \in T)\}$ donde si $T = \mathbb{R}$ se dará el caso unidimensional, pero también puede ser $T = \mathbb{R}^2$ en imágenes, u otras expresiones para casos más complejos. Además, hasta ahora tenemos una única observación, pero para clasificar, por ejemplo, serán necesarios conjuntos de datos. Una *Muestra* o *Dataset Funcional* de dimensión n, $\chi_1, \chi_2, \cdots, \chi_n$ es el conjunto de las observaciones de n variables funcionales $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n$ idénticamente distribuidas con $\mathcal{X}$. En 4, página 49 se verán algunos ejemplos de estos conjuntos de datos y hay buenas secciones dedicadas a presentar algunas de las muestras más comunes en varias referencias [35, 80, 5]. En la próxima sección se hablará brevemente de las dificultades para generarlos, pero lo que más nos interesa es que los avances técnicos permiten obtener esta clase de muestras de manera más eficiente y eficaz, por lo que actualmente es más fácil encontrar datos para trabajar. Esto, junto con el propio interés de los datos, son unas de las razones del auge del FDA en la actualidad, surgiendo una gran cantidad de conjuntos de datos en todos los campos: química,

economía, medicina, seguridad y muchos otros.

Por otra parte son necesarias herramientas para extraer la información de las funciones. Estas herramientas vendrán dadas por el espacio en el que habite la función y determinarán la dificultad del problema. De este modo interesará, o por lo menos facilitará el trabajo, poder estimar una serie de magnitudes. En primer lugar querremos saber si las cosas están lejos o cerca en el espacio de funciones.

Definition 1 *Sea $\mathcal{F}$ un espacio funcional, es un espacio métrico si en $\mathcal{F}$ existe una aplicación distancia $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ que para cualesquiera dos elementos f, g de $\mathcal{F}$ cumple:*

1. $d(f,g) \geq 0 \quad y \quad d(f,g) = 0$ *si, y sólo si* $f = g$
2. $d(f,g) = d(g,f)$ *(Simétrica)*
3. $d(f,g) \leq d(f,h) + d(h,g)$ *(Desigualdad Triangular)*

Otro punto importante para definir un objeto es su tamaño.

Definition 2 *Sea $\mathcal{F}$ un espacio vectorial funcional, es también un espacio normado sobre un cuerpo $\mathbb{K}(= \mathbb{R}$ ó $\mathbb{C})$ si en $\mathcal{F}$ existe una aplicación norma $\| \cdot \|$: $\mathcal{F} \to \mathbb{R}$ que para cualesquiera dos elementos f, g de $\mathcal{F}$ cumple:*

1. $\| f \| \geq 0 \quad y \quad \| f \| = 0$ *si, y sólo si,* $f = 0$
2. *Para todo* $\lambda \in \mathbb{K}$, $\| \lambda f \| = | \lambda | \| f \|$
3. $\| f + g \| \leq \| f \| + \| g \|$ *(Desigualdad Triangular)*

Todo espacio normado se puede convertir en un espacio métrico con la distancia inducida por la norma $d(f,g) = \| f - g \|$. Además, si el espacio es completo (toda sucesión de Cauchy converge con la norma) se dice que es un espacio de Banach. Los espacios de Banach pueden ser un buen punto de partida, pero generalmente se pide un poco más, esto es, conocer las orientaciones, posiciones relativas o ángulos.

Definition 3 *Sea $\mathcal{F}$ un espacio funcional, será un espacio euclídeo sobre un cuerpo $\mathbb{K}$ si en $\mathcal{F}$ existe una aplicación producto escalar $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{K}$ que para cualesquiera dos elementos f, g, h de $\mathcal{F}$ cumple:*

1. $\langle f, f \rangle \geq 0 \quad y \quad \langle f, f \rangle$ *si, y sólo si* $f = 0$
2. *Para todo* $a, b \in \mathbb{K}$ $\langle af + bg, h \rangle = a\langle f, h \rangle + b\langle g, h \rangle$ *(Linealidad por la izquierda) y análogamente linealidad por la derecha.*

3. *Sea* $\langle f, g\rangle = \overline{\langle g, f\rangle}$ *(Hermiticidad)*

Todo espacio euclídeo se convierte en un espacio normado con la norma inducida por el producto escalar, $\| f \| = \sqrt{\langle f, f\rangle}$. Si el espacio además es completo se dice que es un espacio de Hilbert. Generalmente se intenta trabajar en espacios de Hilbert por la comodidad, aunque no siempre es posible. Sirvan como ejemplo los L-espacios definidos como $L_p(S, \mu) = \{f : S \to \mathbb{K} \text{ tal que } f - medible, \| f \|_p^p = \int \mid f \mid^p d\mu < \infty\}$ donde (S, μ) es el espacio de medida. Estos espacios en el caso más usual $L_p[a, b]$ con $\mathbb{K} = \mathbb{R}$ son de Hilbert para $p = 2$ y de Banach para $p \neq 2$. Además de estas consideraciones sobre análisis funcional serán necesarias las herramientas básicas de estadística como son el cálculo de medias, varianzas o correlaciones, otras más avanzadas como las medidas de profundidad, así como algunos resultados y teoremas básicos en el campo a los que se hará referencia en función de su relevancia en el presente trabajo.

Por último, en lo referente a notación se seguirá la nomenclatura adoptada en [35] de modo que $\mathcal{X}$ hará referencia a una variable funcional y χ o f a una observación de la misma. Respectivamente X y x simbolizarán la variable aleatoria y la observación en el caso multivariado. También, hablaremos en general de curvas refiriéndonos a los datos funcionales. Aunque algunos no sean estrictamente curvas no hay pérdida de generalidad ya que todo sería fácilmente adaptable.

2.2. Características y Dificultades

Como ya se ha podido vislumbrar, el utilizar datos funcionales conlleva unas particularidades que hacen que los métodos tradicionales no sirvan o queden cortos. Esto es debido principalmente a tres de sus características: la dimensión, la correlación y el trabajar en espacios funcionales.

La dimensión complica la obtención de información enormemente ya que no podemos trabajar con objetos infinitas. El primer problema lo encontramos en la propia captura de datos ya que no es posible capturar la curva integra. En este sentido los avances técnicos palían esta pérdida de información con rejillas cada vez más finas, si bien muchas veces son necesarias técnicas de interpolación, suavizado u otras para completar los datos, las cuales pueden introducir más error. Una vez capturado el dato, además del procesado propio de cualquier problema (imperfecciones, outliers, etc.), la altísima dimensionalidad en la práctica, e infinita en la teoría, hacen que haya que elegir una representación adecuada para trabajar. El problema de la representación es muy importante en FDA y hay muchísimo material al respecto, en próximas secciones se presentarán los métodos más utilizados con sus ventajas e inconvenientes. Una vez representado el dato se mitiga en parte el inconveniente de la dimensionalidad, aunque al reducirla estamos perdiendo información y saber qué información es la importante (en este caso para clasificar) es un punto de discusión

interesante y origen en parte de este trabajo en el que se intentará ver una forma 'inteligente' de reducir la dimensión de la malla.

Si la dimensión genera problemas en la captura y manipulación de los datos provocando pérdidas de información, la correlación genera importantes problemas en los algoritmos clásicos a la hora de extraer la información. En el caso de datos funcionales los distintos puntos de la curva están altísimamente correlacionados. Esta correlación no indica más que cercanía y habrá muchos puntos muy parecidos, con lo que introducimos redundancia al sistema. Como es bien sabido, la redundancia puede empeorar los resultados de un algoritmo haciendo que no se consideren otros puntos, sobreajustando, e incluso anularlo provocando matrices cuasi-singulares, como ocurre con el discriminante lineal. Por tanto, el FDA necesitará nuevos métodos o cambios sustanciales en algunos antiguos para lidiar con el problema de la alta correlación. En el trabajo se expondrán algunas de las soluciones actuales para este problema y mediante el método de mRMR [28] se tratará de mitigar el efecto de la redundancia en la solución propuesta.

Finalmente, la propia naturaleza de los datos plantea otra dificultad. Dependiendo del espacio en el que vivan las funciones puede que no tengamos ni siquiera una métrica, pero aún teniéndola no es trivial definir el concepto de cercanía o de similitud entre dos funciones, incluso puede que dependiendo del problema nos interese más un criterio u otro. Por todo esto, tienen una gran relevancia en el FDA la elección de la métrica, o semi-métrica como veremos a continuación. Además, en esta misma línea, es complicado establecer cual es el elemento central de un conjunto de funciones, o que observaciones están en el extremo. Recientemente, se han venido proponiendo unas nuevas medidas estadísticas, las medidas de profundidad [24, 66], que pueden ser más útiles para este tipo de datos que los estadísticos clásicos.

2.2.1. Análisis Exploratorio de Datos (Representación)

Al contrario que en el caso multivariado en el que el gráfico de una nube de puntos en $\mathbb{R}^2$ nos aporta mucha información, en el caso funcional un gráfico puede ser muy poco informativo ya que los datos pueden estar sujetos a métricas no usuales y esta representación sólo engañaría nuestra mirada. Uno de los objetivos del FDA es la representación de los datos funcionales de manera que faciliten posteriores análisis y faciliten la apreciación de sus características [81]. Por ello, el primer paso en todo problema con datos funcionales, independientemente del objetivo, es el análisis exploratorio de los datos para poner de manifiesto sus características o prepararlos para posteriores manipulaciones. Esta fase tiene como parte fundamental la representación de los datos, como veremos a continuación, pero también hace referencia a otros aspectos: el registro y alineamiento de los datos es necesario ya que las muestras pueden tomarse, por ejemplo, con

pequeños retrasos, como el caso de una grabación que no empieza siempre en el mismo instante. En estos casos hay que cuadrar las muestras antes de empezar. Las visualizaciones previas de los datos también entran en este epígrafe y presentan sus propias dificultades. En otras ocasiones, la propia curva no nos dice mucho y es útil estudiar sus derivadas y compararlas con la curva original. Todo esto es necesario, tanto por la naturaleza de los datos como por los métodos de captura o los errores en la captación de los mismos. Así, por ejemplo, puede que los instantes de muestreo sean distintos entre las observaciones, o incluso que el intervalo varíe. Además, toda medición introduce un error o ruido, obteniendo en realidad una muestra $y = \chi(t) + \epsilon$. Esto debe considerarse en algún momento del procesado y filtrar la señal en la medida de lo posible para aproximarse al deseado $\chi(t)$. Hay distintas metodologías aplicables, en [81] se puede encontrar una buena muestra. En este documento se considerará que se han tenido en cuenta todos estos detalles para no desviarnos demasiado del asunto principal. Asimismo, en el análisis exploratorio de datos se incluyen los análisis de la variabilidad del conjunto tales como Componentes Principales o Correlaciones Canónicas. Finalmente, esta fase concluiría con las conclusiones sobre los datos o la elección del modelo a utilizar.

Volviendo al problema de la representación, como se ha comentado, los datos funcionales tienen dimensión infinita. Esto hace que puedan ser más o menos manejables en la teoría, pero presenta dificultades, e impide cualquier aproximación computacional en la práctica. Por eso, para atacar los problemas hay pasar los ejemplos a una dimensión finita y, en la medida de lo posible reducida. Éste es un problema inherente a estos datos, y es de capital importancia, ya que una mala conversión puede hacer que se pierda información discriminante. Por tanto, es un punto crítico sobre el que hay una extensísima bibliografía. Si bien en cada artículo se propone un procedimiento de representación, pese a la mayor o menor originalidad o combinación de sistemas los métodos de representación se pueden agrupar en dos grandes familias: la discretización y la elección de una base de funciones reducida. Los sistemas que se engloban bajo estos epígrafes son ampliamente utilizados y no se puede decir que una filosofía sea mejor que otra, en general los autores las eligen según el problema al que se enfrentan. Sin embargo, en la discretización hay una vía poco explorada, que es la que proponemos en este trabajo. Haciendo una discretización muy fina, y eligiendo cuidadosamente los puntos que utilizaremos, parece razonable pensar que se obtendrían buenos resultados. Así trabajaría una selección de variables, y en está línea hay algunos trabajos, si bien escasos, como [61] donde se plantean unas ventanas con un cierto suavizado, o la aparición de métodos que podríamos llamar en un sentido amplio de selección de variables en este último año [92].

2.2.1.1. Discretización

Como para todo problema básico, para éste se han propuesto distintas soluciones. La primera que se nos ocurre (y posiblemente la más utilizada debido a su simplicidad) es la discretización. Como ya hemos dicho, los datos funcionales se toman en función de un continuo. Podemos suponer sin pérdida de generalidad que ese continuo es el tiempo t, y que tenemos un dato funcional $f(t)$ que tiene valores en el intervalo $[t_{\text{mín}}, t_{\text{máx}}]$. La discretización más sencilla consiste en tomar una partición del tiempo $\{t_i\}_{i=0}^{n}$ tal que $t_{\text{mín}} \leq t_0 < t_1 < \cdots < t_n \leq t_{\text{máx}}$ y tomar como atributos los valores $\{f(t_i)\}_{i=0}^{n}$. Sin embargo, aun en el caso más sencillo tenemos que tener en cuenta múltiples factores:

- Hay que elegir el número de elementos de la partición n, ya que si la partición es demasiado gruesa nos dejaremos por fuerza información discriminante. Sin embargo, una partición muy fina generará variables muy redundantes (con los problemas que ello acarrea en la clasificación) y elevará el coste computacional.

- También puede ser conveniente definir la separación entre los elementos de la partición. La selección trivial es una partición equiespaciada en la que los $n+1$ puntos disten $\frac{t_{\text{máx}}-t_{\text{mín}}}{n}$, pero no se tiene en cuenta la forma de la función. Esto hace que no sea una decisión óptima ni en lo referente a la eficacia (puede que no se tomen los puntos más discriminante por caer entre medias) como de la eficiencia (para mitigar el efecto anterior se pueden tomar más puntos de lo necesario). Otra posibilidad es realizar una partición estocástica, lo que aumenta el primero de los riesgos y para arreglarlo habría que promediar o usar técnicas de Monte Carlo que elevarían mucho el coste. Lo ideal sería estudiar el problema y elegir la partición en función de la morfología de los datos eligiendo pocos puntos en las zonas con poca capacidad discriminante y haciendo una partición más fina en las más discriminantes.

- Además es preciso decidir el criterio por el que se seleccionan los valores para los elementos de la partición. La opción más sencilla es tomar el valor de la función en el punto medio de cada intervalo de la rejilla correspondiente, pero de nuevo, esto obvia parte de la información de la función (que puede útil). Para esto también hay multitud de soluciones con mayor o menor complejidad como interpolación, splines, etc. [81].

Una posible solución para no perder la información de la función con la discretización y poder trabajar de manera eficiente con los datos resultantes sería hacer una discretización muy fina (mantenemos "todos" los puntos) y después reducir este conjunto de variables mediante algún algoritmo inteligente.

2.2.1.2. Representación en una Base

La otra opción más utilizada para representar funciones es dar las coordenadas de su proyección en algún sub-espacio funcional de dimensión finita. Esto se hace generalmente considerando su desarrollo en una cierta base prefijada (por ejemplo una base trigonométrica) y truncando este desarrollo para quedarnos sólo con un número finito de términos.. De esta manera, los datos siguen siendo funciones (una composición) y por tanto conservan ciertas propiedades, aunque al ser aproximaciones los resultados se verán influidos por el tipo de base elegida. Hay muchas formas de hacerlo, pero en general, si definimos una base de funciones como el conjunto de funciones independientes $\{\phi_k\}_{k\in\mathbf{N}}$ tales que cualquier función puede aproximarse tan bien como se quiera mediante una combinación lineal de ellas, se trunca en un cierto K con un error asumible, es decir, $\mathcal{X}(t) = \sum_{k\in\mathbb{N}} c_k\phi_k(t) \sim \sum_{k=1}^{K} c_k\phi_k(t)$, donde los c_i son los coeficientes en la nueva base. De nuevo, la dificultad de esta operación vendrá determinada por el espacio en el que habiten las funciones. Por ejemplo, en todo espacio de Hilbert existe una base ortonormal tal que $\mathcal{X}(t) = \sum_{k=1}^{\infty}\langle\mathcal{X}, e_i\rangle e_i$, y teniendo un producto interno es fácil calcular los coeficientes resolviendo el siguiente sistema:

$$\begin{pmatrix} \langle\mathcal{X},\phi_1\rangle \\ \vdots \\ \langle\mathcal{X},\phi_i\rangle \\ \vdots \\ \langle\mathcal{X},\phi_K\rangle \end{pmatrix} = \begin{pmatrix} \langle\phi_1,\phi_1\rangle & \cdots & \langle\phi_i,\phi_1\rangle & \cdots & \langle\phi_K,\phi_1\rangle \\ \vdots & & \vdots & & \vdots \\ \langle\phi_1,\phi_i\rangle & \cdots & \langle\phi_i,\phi_i\rangle & \cdots & \langle\phi_K,\phi_i\rangle \\ \vdots & & \vdots & & \vdots \\ \langle\phi_1,\phi_K\rangle & \cdots & \langle\phi_i,\phi_K\rangle & \cdots & \langle\phi_K,\phi_K\rangle \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ \vdots \\ c_i \\ \vdots \\ c_K \end{pmatrix}$$

Así se tienen observaciones en una base finita, lo que además, si los elementos de la base son funciones diferenciables, facilita mucho el cálculo de derivadas ya que $\mathcal{X}^{(p)}(t) = \sum_{k\in\mathbb{N}} c_k\phi_k^{(p)}(t) \sim \sum_{k=1}^{K} c_k\phi_k^{(p)}(t)$

En cualquier caso, salvo para ejemplos muy sencillos, la expresión de la función seguirá siendo infinita o demasiado grande, por lo que habrá que truncar la serie. Aquí entra en juego la elección de K, que se comporta como otro parámetro del modelo midiendo el grado de interpolación/suavizado de la función, cuanto menor sea el número de elementos en la base menor será el coste computacional pero se puede realizar un suavizado excesivo perdiendo información relevante. Elegir el K en el que finalizar es a menudo punto de discusión y no hay un método claro para hacerlo, se suele elegir con algún mecanismo de validación.

Además hay muchas bases con distintas propiedades en función de los elementos que la compongan. Algunos ejemplos son exponenciales, potencias polinomios. No hay que cometer el error de pensar que al tomar una base se reduce el problema al caso multivariado, ya que los resultados dependerán enormemente de la elección de las ϕ_i. Unas bases se ajustarán mejor a los datos que otras, de modo que eligiendo bien la naturaleza de estas funciones podemos reducir la K y tener mejores descriptores. A continuación hablaremos brevemente de las tres más utilizadas [81]:

Fourier La base de Fourier es una de las más antiguas y conocidas junto con la de los polinomios. Mientras la segunda ha quedado relegada a los problemas más sencillos, la primera se sigue utilizando en múltiples problemas [10], pero aunque puede utilizarse en casi todos los casos, no suele obtener resultados espectaculares.

La extensión de Fourier de $\chi(t)$ es de la forma

$$\hat{\chi}(t) = c_0\phi_0 + \sum_r c_{2r-1}\phi_{2r-1}(t) + c_{2r}\phi_{2r}(t)$$

donde $\phi_0(t) = \frac{1}{\sqrt{T}}$, $\phi_{2r-1}(t) = \frac{sen(r\omega t)}{\sqrt{T/2}}$, $\phi_{2r}(t) = \frac{cos(r\omega t)}{\sqrt{T/2}}$ forman una base periódica de periodo $T = \frac{2\pi}{\omega}$ que será ortonormal si los $\{t_j\}$ se toman equiespaciados en $[0, T]$.

La base de Fourier ha sido tradicionalmente utilizada para series temporales largas debido a que la transformada rápida de Fourier (FFT) [78] permite calcular todos estos coeficientes de manera eficiente (en $O(n \log n)$ operaciones) cuando el número de puntos n es potencia de 2 y están equiespaciados. Sin embargo, en la actualidad, las técnicas para B-splines y wavelets igualan o superan esta eficiencia computacional.

Otras de sus características son la fácil diferenciabilidad y la periodicidad. Debido a sus características (está formada por senos y cosenos) es muy sencillo conseguir la expresión derivada en cualquier orden, y al ser una base de funciones periódicas, se ajustará mejor a problemas en los que los datos muestren una cierta periodicidad.

En resumen, la expansión de Fourier es bien conocida, así como sus limitaciones. Esta representación será especialmente útil para funciones estables, sin grandes variaciones y con una curvatura más o menos constante. También será conveniente que los datos muestren una cierta periodicidad, por ejemplo, la variación anual de temperatura o precipitaciones. Fourier genera expansiones suaves en el sentido de diferenciabilidad, por lo que no será apropiada para funciones que tengan discontinuidades o que las presenten en sus derivadas de orden bajo.

B-Splines Puede que los splines sean la aproximación más utilizada en el caso de datos no periódicos reemplazando de alguna manera a los polinomios que quedan contenidos en ellos. Los splines combinan la eficiencia computacional de los polinomios con una mayor flexibilidad, que muchas veces hace que la K necesaria para obtener buenos resultados sea pequeña. Parte del éxito de estos métodos es que se han desarrollados sistemas para funciones de spline con un coste computacional del orden de n, lo que los hace interesantes para grandes cantidades de datos.

El primer paso para definir un spline es dividir el intervalo $T = [a, b]$ en L subintervalos separados por los puntos $a = t_0, t_1, \cdots, t_L = b$. En cada uno de estos intervalos, el spline es un polinomio de un cierto orden m que trata de ajustar la curva. Se entiende como el orden del polinomio m al número de coeficientes que hacen falta para definirlo, es decir, uno más que el grado, la mayor potencia. Estos polinomios deben coincidir en los nodos de la partición hasta la

derivada de orden $m-2$, por lo que salvo para polinomios de orden 2 (rectas) las uniones serán suaves en el sentido de diferenciabilidad, como se puede apreciar en la figura 2.2.1. También existe la posibilidad de reducir esta diferenciabilidad introduciendo más nodos en un mismo punto como se explica en [14], con el objeto de ajustarse mejor a funciones con discontinuidades o ciertos puntos críticos. Con estas restricciones el número de grados de libertad del spline (los parámetros a definir) será $m+L-1$. Por tanto, un punto importante será definir el orden de los polinomios, el número de particiones y el punto en que se fijen los nodos. Normalmente, a mayor grado del polinomio y a particiones más finas, la función de spline se ajusta mejor, pero el coste es mayor y no siempre se consigue un mejor resultado. En el caso más simple, si no hay nodos interiores, el spline se convierte en un polinomio con m grados de libertad.

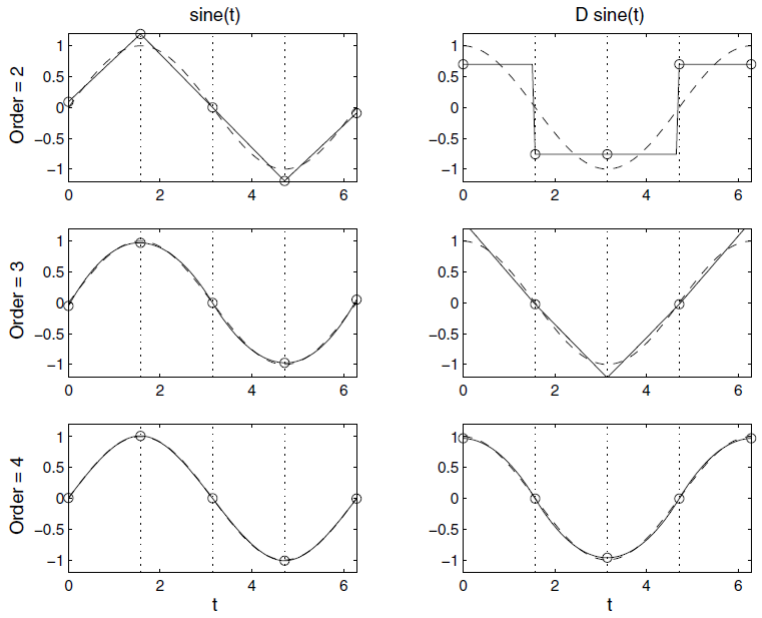

Figura 2.2.1: En la parte izquierda, la línea continua representa la función spline de un cierto orden y como se ajusta a la función seno marcada por la traza discontinua. En la derecha se muestra el ajuste de la respectiva derivada del spline a la función coseno. Las líneas verticales punteadas indican los nodos.

Las llamadas bases de B-splines permiten aproximar todas estas funciones facilitando así su manejo. Schoenberg acuñó el término B-spline para referirse a estas bases como abreviatura de splines básico. Sus características se basan en las siguiente propiedades:

- Cada elemento de la base $\phi_k(t)$ será una función spline de orden m y partición τ.
- Cualquier combinación lineal de funciones spline es una función spline.

- Cualquier función spline de orden m sobre la partición τ se puede expresar como combinación lineal de las funciones de la base.

Además, las bases de splines tienen la particularidad de que los elementos de un B-spline cumplen la propiedad de soporte compacto, de modo que si la base es de orden m sus funciones son distintas de cero (y positivas) en un máximo de m subintervalos adyacentes. Esto tiene ventajas computacionales al obtenerse una matriz de cambio de base banda, de modo que no presenta grandes desventajas frente a bases ortogonales como Fourier o Wavelets.

Si tomamos la notación usual $B_k(t, \tau)$ para representar el valor del k-ésimo elemento de la base sobre la partición τ en el instante t, la función spline $S(t)$ se representa:

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

Esta base se puede calcular fácilmente y de forma numéricamente estable con el algoritmo de Boor [14]. Aunque hay otros sistemas este es el más utilizado y está implementado en múltiples lenguajes como MATLAB o R.

Wavelets La transformada wavelet, consistente en la representación de funciones mediante ondas, es el más reciente de estos métodos de representación, ya que aunque los primeros trabajos empezaron con Haar a comienzos del siglo XX, la transformada continua no se formuló hasta 1975 (Zweig) y fueron bautizados con los trabajos de Morlet y Grossmann a principios de los 80. Los wavelets u ondículas son un sistema de representación muy extendido en el manejo de señales, utilizándose la transformada discreta para la codificación de señales y la continua en el análisis de señal, pero también están ganando terreno en ámbitos como la compresión de datos, la sismología o la genética, y también en la clasificación funcional [8].

La trascendencia de estás transformaciones se debe a que combinan el análisis de frecuencias de las series de Fourier con la información temporal (o espacial). Además, una base de wavelets se puede adaptar fácilmente para trabajar con funciones discontinuas o no diferenciables, al contrario que Fourier, y en el caso de la transformada discreta (DWT) [78] se pueden obtener los coeficientes en $O(n)$, mejorando el $O(n \log n)$ de la FFT.

La idea de los wavelets aprovecha que cualquier función de L^2 puede representarse mediante una función ψ apropiada, llamada wavelet madre, y considerando todas sus traslaciones y dilataciones de la forma $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$, con j y k enteros. Con este resultado, para generar la base de wavelets se toma un wavelet madre con soporte compacto (algunos ejemplos en 2.2.2 y el llamado wavelet padre, ϕ ortogonal al madre por traslación y cambio de escala. La expansión de una función f, se basa en el llamado análisis multiresolución, la observación de señales a distintas escalas de resolución, en el sentido de que el coeficiente ψ_{jk} aporta información sobre f cerca de

la posición $2^{-j}k$ en una escala 2^{-j}. Una vez elegida la base, es decir, fijado el $j = J$ se tiene que

$$f(t) = \sum_k c_{J,k}\phi_{J,k}(t) + \sum_{j=J}^{\infty}\sum_k d_{j,k}\psi_{j,k}(t)$$

donde $c_{J,k}$ es el coeficiente de aproximación o señal suave

$$c_{J,k} = \langle f, \phi_{J,k}\rangle = \int f(t)\phi_{J,k}(t)dt$$

y $d_{j,k}$ es el coeficiente de detalle

$$d_{j,k} = \langle f, \psi_{j,k}\rangle = \int f(t)\psi_{j,k}(t)dt.$$

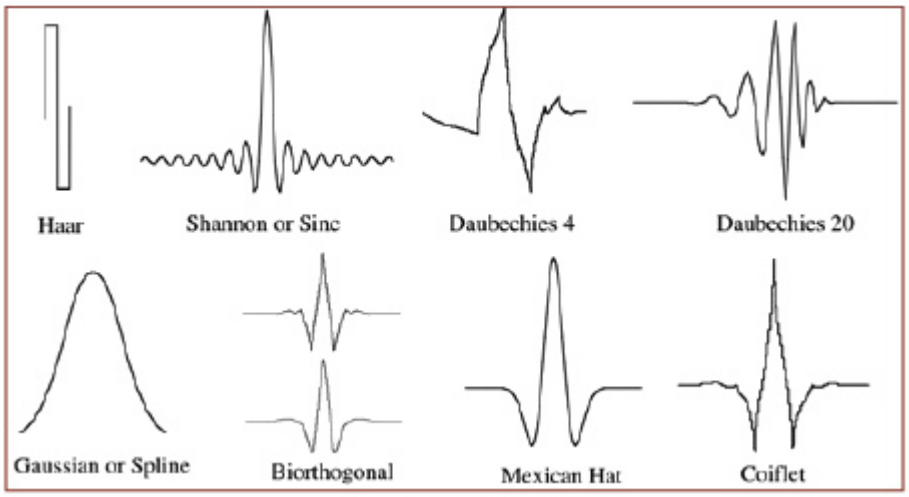

Figura 2.2.2: Distintos ejemplos de wavelets madre.

2.2.2. La noción de cercanía

Disponer de una cierta medida de cercanía mutua de los datos es un elemento fundamental para el análisis estadístico y, en particular, para analizar nociones básicas como la dispersión. Obviamente, las nociones de cercanía o de dispersión de los datos están totalmente condicionadas por la forma en la que se midan las distancias. Así, la primera preocupación para trabajar con datos funcionales será determinar el espacio funcional en el que se trabajará. Esto determinará las herramientas disponibles posteriormente. Por tanto, la métrica del espacio funcional que se elija debe ser coherente con los datos que se describan. De este modo habrá que tener en cuenta si la información descriptiva está en la escala, en la forma, etc. Por ejemplo, si se piensa que el cambio de

escala es informativo se puede tomar L^2 como espacio de referencia, pero si la información está en la curvatura puede que una semi-norma basada en derivadas nos sea más útil. En la bibliografía se suele utilizar $L^2[a, b]$ por su generalidad y sus buenas propiedades, y así lo encontraremos en el libro de Ramsay y Silverman [81]. En otras referencias como la de Ferraty y Vieu [35] se realiza un planteamiento más general considerando espacios funcionales con distintas métricas o semi-métricas que en algunos casos pueden ser más apropiados.

La medida de la proximidad entre objetos juega un papel fundamental en toda la estadística. En el caso finito, típicamente $\mathbb{R}^d$, hay una equivalencia entre todas las normas en el sentido de que definen la misma métrica, y por tanto, la misma topología. Por esto, la elección de una u otra puede tener menos relevancia, aunque repercutirá en aspectos computacionales y puede tener alguna implicación más importante, incluso decisiva, dependiendo del tipo de problema al que nos enfrentemos (por ejemplo, la norma euclídea usual tiene asociado un producto escalar con el que medir ángulos). Sin embargo, en el caso funcional no se da si quiera esta equivalencia y los problemas serán atacados de forma distinta en función de la métrica que elijamos, lo que convierte la elección de dicha métrica en una decisión importante. En este sentido, puede que incluso la elección de cierto espacio métrico sea insuficiente y como se propone en [35], sea ventajoso el uso de espacios semi-métricos.

Definition 4 *Sea $\mathcal{F}$ un espacio funcional, es un espacio semi-métrico si en $\mathcal{F}$ existe una aplicación semi-métrica $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ que cumple:*

1. *Para todo $f \in \mathcal{F}$, $d(f, f) = 0$.*
2. *Para todo f, g, $h \in \mathcal{F}^3$, $d(f, g) \leq d(f, h) + d(h, g)$*

Se observa que una semi-métrica es una métrica, excepto que $d(f, g) = 0$ no implica $f = g$, es decir, puede haber funciones en un mismo sitio y no ser iguales. Análogamente se puede definir una semi-norma y un espacio semi-normado con las mismas propiedades que la norma (2, página 7) haciendo que elementos no nulos puedan tener norma cero, $\| f \| = 0 \not\Rightarrow f = 0$. Estás restricciones de menos son las que confieren a las semi-métricas una mayor flexibilidad. En el capítulo 3 de [35] se describen tres familias de semi-métricas adaptadas a datos funcionales:

- **Basadas en FPCA.** El FPCA es la adaptación del Análisis de Componentes Principales al caso funcional y se comentará en 2.3.1, página 22. Dada una variable aleatoria funcional $\mathcal{X}$, FPCA nos proporcional la siguiente expansión

$$\mathcal{X} = \sum_{k=1}^{\infty} \left(\int \mathcal{X}(t) v_k(t) dt \right) v_k,$$

donde los v_k son los autovectores de la matriz de covarianzas ordenados decrecientemente por autovalor. Truncando esta expansión hasta dimensión q se puede definir la familia de semi-métricas

$$d_q^{PCA}(\mathcal{X}_i, \chi) = \sqrt{\sum_{k=1}^{q} \left(\int [\mathcal{X}_i(t) - \chi(t)] v_k(t) dt \right)^2}$$

Estas semi-métricas sólo pueden utilizarse con datos medidos en los mismos puntos y tomados de una partición suficientemente fina para que los estimadores empíricos sean consistentes. En cambio, permiten abordar funciones irregulares.

- **Basadas en PLS.** Pensada para situaciones en las que se observa una variable respuesta con p "clases". El tener en cuenta esta respuesta hace de estas semi-métricas una herramienta más adecuadas que las basadas en FPCA para muchos problemas de regresión y clasificación. Si los v_k son los vectores de proyección obtenidos de PLS como se explica en 3.1, página 38 y aproximando la integral numéricamente con los w_j como pesos de cuadratura [35] la expresión empírica de esta semi-métrica es:

$$d^{PLS}(x_i, x_{i'}) = \sqrt{\sum_{k=1}^{p} \left(\sum_{j=2}^{J} w_j (\chi_i(t_j) - \chi_{i'}(t_j)) v_k(t) dt \right)^2}.$$

  Estas semi-métricas tienen las mismas propiedades que las basadas en FPCA.

- **Basadas en derivadas.** Al contrario que las anteriores, las semi-métricas basadas en derivadas sólo pueden aplicarse a funciones suaves, ya que toman las distancias entre derivadas,

$$d_q^{deriv}(\chi_i, \chi_{i'}) = \sqrt{\int \left(\chi_i^{(q)}(t) - \chi_{i'}^{(q)}(t) \right)^2 dt},$$

  donde q es el orden de derivación. Si se hace $q = 0$ se tiene la norma usual de L^2. Para simplificar los cálculos de las derivadas será útil representar la función en una base, y haciéndolo así no estaríamos sujetos a una rejilla fija. Estás semi-métricas son apropiadas cuando la información a estudiar esté en la curvatura, y suelen requerir un trabajo previo para la correcta estimación de las derivadas (generalmente no son observables), frecuentemente suavizando con kernels o splines.

2.2.3. Medidas de Profundidad

Para estudiar, y en concreto para clasificar cualquier tipo de datos es importante poder representarlos correctamente (2.2.1, página 9) y tener una idea de sus posiciones relativas (2.2.2, página 16). Pero también es fundamental saber cuáles están en los extremos y qué elementos son los centrales, en definitiva, poder ordenar las funciones. Esto no es trivial, y para cubrir esta necesidad surge la idea de la profundidad.

La noción de profundidad fue desarrollada en primer lugar en el contexto multivariado para generar estadísticos de orden y listas ordenadas en altas dimensiones. Dada una distribución de probabilidad F en $\mathbb{R}^d$, la profundidad asigna a cada punto x un valor positivo $D(x, F)$. Desde los primeros estudios de Mahalanobis en 1936, varios autores han propuesto y analizado distintas medidas de profundidad para datos multivariados. Es posible encontrar información y referencias en [113]. Se podría pensar en la generalización de estas medidas a los datos funcionales, sin embargo, esto puede conducir a medidas computacionalmente intratables o con malas propiedades. Así, en la última década han surgido formulaciones de medidas de profundidad específicas para datos funcionales: [96] proponen una extensión de la profundidad L_1, en [40] se plantea una medida basada en la integración de profundidades univariadas, [24] dan otra vuelta a estas ideas y añaden un nuevo enfoque partiendo de proyecciones aleatorias, y [66] realizan una propuesta gráfica de bandas sobre las representaciones de curvas. A continuación se explican algunas de las versiones de profundidad que pueden ser herramientas para definir estimadores robustos, ordenar o incluso clasificar en función de la profundidad del elemento en las poblaciones entre las que decidir.

Fraiman y Muniz Basada en la naturalidad con que surgen estadísticos de orden en una dimensión, se propone una profundidad integral. Sin pérdida de generalidad se toman funciones definidas en el intervalo $[0, 1]$, entonces, para todo $t \in [0, 1]$, si $F_{n,t}$ es la función de distribución empírica de $x_1(t), \cdots, x_n(t)$, se denota la profundidad univariada de cada dato $x_i(t)$ por $D_i(t) = 1 - \mid \frac{1}{2} - F_{n,t}(x_i(t)) \mid$ y para cada i se define

$$I_i = \int_0^1 D_i(t)dt.$$

En función de este índice de profundidad I_i se pueden ordenar los elementos. Fue propuesto en 2001 por Fraiman y Muniz [40] como una herramienta para definir estimadores y puede adaptarse fácilmente al caso multivariado sustituyendo la integral por el sumatorio apropiado.

Moda h Esta medida observa relaciones de cercanía entre las funciones para asignar la mayor profundidad al elemento que esté de alguna manera "más rodeado". Según este método, la profundidad h de un dato x viene dada por la función

$$f_h(x) = E(K_h(\parallel x - X \parallel))$$

donde X es el elemento aleatorio que describe la población, $\|.\|$ es alguna norma (en el caso funcional puede ser, por ejemplo, la norma L^2), $K_h(t)$ es un núcleo de re-escalado del tipo $K_h(t) = \frac{1}{h}K(\frac{t}{h})$ (en [24] se utiliza el núcleo gaussiano $K(t) = \frac{1}{\sqrt{2\pi}} \exp \frac{-t^2}{2}$ ()) y h es un parámetro de ajuste fijado. Así, se define la moda h de X como el valor más profundo de la muestra obtenido

maximizando la ecuación anterior en x. Finalmente, dado un conjunto aleatorio $X_1, \cdots, X_n$ de X, la versión de la profundidad h queda expresada con la media en lugar de la esperanza

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(\parallel x - X_i \parallel).$$

Proyecciones Aleatorias (RP) Está basado en los trabajos de Cuesta-Albertos, Fraiman y Ransford que introdujeron la idea de las proyecciones aleatorias para funciones al conseguir una extensión del teorema de Cramer-Wold, que caracteriza una medida de probabilidad en términos de las proyecciones unidimensionales, aplicable a espacios de Hilbert [22]. A partir de esta idea están surgiendo muchas aplicaciones estadísticas de las que se pueden encontrar referencias en [71]. En particular, en [24] se plantea una medida de profundidad en función de las medidas obtenidas en varias proyecciones unidimensionales aleatorias de los datos.

Dado un conjunto $X_1, \cdots, X_n$, se toma una dirección aleatoria a independiente de los X_i y se proyecta sobre ella. Después, se define la profundidad de X_i con alguna medida de profundidad simple unidimensional. En el caso de datos funcionales se puede asumir que nos encontramos en $L^2[0,1]$ y utilizar su producto interno para proyectar. En el caso multivariado se proyecta con el producto escalar estándar. Una vez se tiene este estimador de profundidad se pueden tomar otras direcciones y promediar los resultados.

2.3. Problemas más importantes

Según [81] los objetivos del análisis de datos funcionales son:

- Representar los datos de forma que se facilite su estudio y se pongan de manifiesto sus características.
- Estudiar la variabilidad entre los datos.
- Explicar la variación de una salida dependiente a partir de una entrada independiente.
- Comparar distintos conjuntos de datos y los elementos que los componen respecto a ciertos criterios.

Para alcanzar estos objetivos surge la necesidad de crear metodologías. Si bien ya se ha hecho referencia a alguno de estos objetivos y mecanismos asociados al hablar del análisis exploratorio de datos (2.2.1, página 9), el resto da pie a los problemas más importantes, y que acaparan la mayor atención en el FDA. Así, para la representación y puesta en valor de características ocultas surge lo que algunos autores llaman análisis factorial, cuyo mayor exponente es el Análisis de Componentes Principales Funcional. El tercer punto motiva el estudio de la regresión, y el último remite

a los problemas de clasificación en sus dos vertientes: asignar elementos a un grupo determinado y formar subgrupos de un conjunto mayor.

A continuación se describirán brevemente dichos problemas y se comentarán algunas de las principales soluciones y tendencias actuales. El epígrafe referente a la clasificación supervisada será más completo y extenso al ser el centro de atención de este trabajo. Para tener una noción general, se puede encontrar bastante información y referencias en [81] y [35]. El primero presta especial atención al problema de la regresión y dedica un capítulo a FPCA, mientras que en el segundo se utilizan métodos no-paramétricos para regresión y clasificación. Para ver las últimas aportaciones, además de los artículos específicos, son recomendables las revisiones de Manteiga y Vieu [68] y Febrero [32], esta última centrada en la regresión, así como [5] para los temas referentes a clasificación. Todos estos documentos han sido la base de las siguientes secciones, pero antes de pasar a ver los problemas concretos, y debido al interés que están generando recientemente los métodos no-paramétricos en algunos sectores del FDA (por ejemplo en el grupo STAPH) conviene hacer una pequeña mención a este enfoque.

La importancia de la estadística no-paramétrica radica en que resuelve algunos de los problemas de la estadística clásica para datos funcionales de forma "natural". Por ejemplo, los métodos tradicionales no trabajan bien con datos tan correlados como los funcionales (matrices de covarianzas cuasi singulares) o con un ratio tan extremo de variables sobre el número de patrones. Por tanto, es necesario desarrollar unos métodos que trabajen bien con la estructura de las funciones. La mayoría de los métodos aplican un modelado para los objetos a estimar o para sus distribuciones, siendo poco generalizables o incurriendo en imprecisiones para poder generalizarlos. Por otro lado, los métodos no-paramétricos no asumen ninguna hipótesis sobre los datos ni sus distribuciones, lo que los hace mucho más generales. Formalizando estas ideas se fijan las diferencias entre métodos paramétricos y no-paramétricos, y como se expanden a clasificación funcional [35]:

- Sea $X \in \mathbb{R}^d$ una variable aleatoria, y sea ϕ una función definida en $\mathbb{R}^d$ y dependiente de la distribución de X. Un modelo para la estimación de ϕ consiste en introducir un conjunto de restricciones de la forma $\phi \in C$. Este modelo se llama modelo paramétrico para la estimación de ϕ si C tiene dimensión finita. En otro caso, el modelo se llama no-paramétrico.

- Sea $Z \in \mathcal{F}$ una variable aleatoria infinita y sea ϕ una aplicación definida en $\mathcal{F}$ y dependiente de la distribución de Z. Un modelo para la estimación de ϕ consiste en introducir un conjunto de restricciones de la forma $\phi \in C$. Este modelo se llama modelo paramétrico funcional para la estimación de ϕ si C tiene dimensión finita. En otro caso, el modelo se llama no-paramétrico funcional.

2.3.1. Análisis de Componentes Principales Funcional (FPCA)

Ligado al análisis exploratorio de datos, el objetivo principal de la extensión funcional del conocido método de Análisis de Componentes Principales es, al igual que en el caso multivariado, dar una representación de los datos mediante el criterio de conservación de la máxima varianza en una dimensión menor de forma que se pongan de manifiesto características latentes de los datos en crudo. Esta representación puede utilizarse únicamente para la visualización o estudio preliminar de los datos, pero a menudo es una herramienta para otros procesos posteriores como la clasificación o la detección de outliers. Por estos motivos, el análisis de componentes principales fue uno de los primeros métodos adaptados al FDA, habiendo gran número de trabajos desde la década de los 50 estudiando sus propiedades y extendiendo sus aplicaciones. Como referencia se puede tomar [81], cuyo octavo capítulo está enteramente dedicado a FPCA, o más recientemente el artículo de Benke, Härdle y Kneip [7] que antes de pasar a temas de funciones de volatilidad hacen un buen resumen de la técnica general.

La extensión de Karhunen-Loève es la herramienta básica para el desarrollo de FPCA. Para aplicarla nos situamos en L^2 con su norma usual, y denotamos por λ_i los autovalores del operador de covarianza Γ, y v_i sus correspondientes autovectores, ordenados ambos de manera descendente. Estos autovectores formas una base ortonormal que permite la expresión de las funciones como

$$\mathcal{X}_i = \sum_{k=1}^{\infty} \beta_{ki} v_k dt$$

donde los coeficientes $\beta_{ki} = \int \mathcal{X}(t) v_k(t) dt$ siendo $E[\beta_{ki}] = \lambda_k$. En casos reales se toman los estimadores empíricos calculando los autovalores y autovectores del operador de covarianza empírico $\hat{\Gamma}$ de $\chi_1, \cdots, \chi_n$ y se trunca la expansión en el número de elementos K que sea conveniente.

Si bien las componentes sintéticas no son fáciles de interpretar, y lo que se captura es la variabilidad global de todo el conjunto, lo cual puede no ser lo más conveniente para algunas aplicaciones posteriores como la discriminación; las componentes principales funcionales son muy utilizadas en muchas aplicaciones importantes debido a que el número de componentes principales necesarias para representar casi toda la varianza es a menudo muy pequeño y por tanto, FPCA supone una forma eficaz y eficiente de acceder a una estructura oculta en primer término. Así, además de los trabajos sobre el propio método y sus características existen multitud de aplicaciones como la definición de semi-métricas y detección de outliers (véase 2.2.2, página 16), un primer paso previo a la clasificación [106] o para facilitar la representación de árboles de regresión [69] y el análisis de los residuos de un modelo de regresión [19].

2.3.2. Regresión

El problema de la regresión ha suscitado siempre mucho interés. Se pueden distinguir distintos tipos de regresión en función de las características de las variables regresoras y de la variable respuesta, pero el problema siempre es el mismo, se trata de predecir o modelar la respuesta a partir de las variables regresoras. De este modo la regresión se aplica en multitud de áreas, desde la economía a la meteorología, e incluso hay autores que prefieren ver la clasificación supervisada como una regresión en la que la variable dependiente es la clase.

En el ámbito funcional, una primera referencia importante es [81] que ya en su edición de 1997 dedica numerosos capítulos a distintas variantes del problema desde un enfoque práctico, dando importancia, por ejemplo, a las propiedades de los estimadores y a cómo penalizar la falta de suavidad. Quizá el caso más estudiado y sobre el que más variantes se han aplicado sea el modelo funcional lineal, con variable regresora funcional y variable respuesta escalar en $L^2[C]$:

$$Y = \alpha_0 + \int_C \alpha(t)\mathcal{X}(t)dt + e(t).$$

En este marco se encuentran trabajos como los de Cardot et al [16] haciendo consideraciones del modelo y teniendo en cuenta los efectos del ruido, los de Hall y Horowitz que en [46] proponen métodos para el ajuste de parámetros en una variante del modelo, o la extensión del modelo para que tenga en cuenta su historia en [48]. Una extensión muy común del modelo consiste en considerar una función r suave general que no esté restringida por un operador lineal. De esta forma se tiene el modelo funcional no-paramétrico, $Y = r(\mathcal{X}(t)) + e(t)$. Esta aproximación ha sido estudiada desde hace años por Ferraty y Vieu entre otros (por ejemplo Preda propone un modelo con kernel reproductores [76]) con múltiples publicaciones cuyas conclusiones se agrupan en la segunda parte de [35]. Allí se definen las adaptaciones funcionales de la esperanza, la mediana y la moda condicionales para utilizarlos en predicción. Todo ello se completa con el marco teórico apropiado y consideraciones prácticas y de implementación.

Cuando la variable respuesta también es funcional se propone un modelo

$$Y(t) = \alpha(t) + \int_S \mathcal{X}(s)\beta(s,t)ds + e(t).$$

Resuelto en [81] representando las funciones en bases reducidas como se explica en 2.2.1, página 9. Actualmente se siguen estudiando los requisitos de convergencia para los estimadores de este modelo, en esta dirección se encuentran trabajos como [23]. Al mismo tiempo se realizan nuevas aproximaciones como las de Hlubinka y Prchal [51] y Chiou y Müller [19] con aplicaciones bioestadísticas y climatológicas respectivamente.

Además existen otras muchas vías de investigación entre las que están, por ejemplo, los trabajos de Escabias et al [31] extendiendo las ideas funcionales para trabajar con respuestas binarias desarrollando un nuevo método de regresión logística funcional, o lo referente a predicción de series temporales, que no despierta tanto interés en el FDA como en otras disciplinas cercanas a la economía.

2.3.3. Clasificación

La última de las grandes áreas del FDA es la clasificación. Independientemente de las divisiones entre métodos paramétricos y no-paramétricos, los distintos tipos de algoritmos o aproximaciones, la clasificación engloba dos grandes ramas conocidas como clasificación supervisada y clasificación no supervisada.

El principal objetivo de la clasificación no supervisada, también llamada clustering, es dividir un conjunto de datos (típicamente grande) $x_1, \cdots, x_n$ en un cierto número de clases k definido, de forma que los miembros de cada clase o cluster tengan algún tipo de *similitud*. Esta similitud vendrá definida por el algoritmo y la métrica utilizados. Uno de los principales retos en clasificación no supervisada es la elección del número de clases, aunque algunos algoritmos aportan sugerencias para esta elección. El clustering se suele utilizar cuando se sospecha de una división oculta en subgrupos de un conjunto de datos y está encaminado a revelar dicha estructura y facilitar la comprensión del problema.

Por su parte, la clasificación supervisada se aplica a problemas en los que k poblaciones (o clases) $P_1, \ldots, P_k$ vienen dadas y bien definidas. En este marco, los individuos son variables aleatorias X, y el conjunto de datos disponible es el conjunto de entrenamiento. Éste se define como el conjunto de duplas $\{(X_i, Y_i), 1 \leq i \leq n\}$, donde los X_i's son observaciones independientes de X, y donde Y_i es la etiqueta que indica la población a la que pertenece la i-ésima observación. El término supervisado hace referencia a que se supone que los elementos del conjunto de entrenamiento están clasificados sin error mediante algún procedimiento no estadístico. Para que el problema tenga solución se asume que la distribución de X en cada población $(X|Y = j)$ es distinta. El objetivo del clasificador será, una vez entrenado, asignar una nueva observación de X a la población adecuada prediciendo Y. Sin pérdida de generalidad nos centraremos generalmente en el caso binario ($k = 2$), de especial importancia en la práctica, por ejemplo clasificación entre sano y enfermo, fraude y no fraude o éxito y fracaso. En cualquier caso, los métodos son fácilmente generalizables a problemas multiclase (se harán algunas referencias) y se ha experimentado con conjuntos de más de dos clases.

Tanto en un caso como en el otro existe abundante literatura y los principales programas tienen paquetes estándar para trabajar con los algoritmos más comunes. En el análisis multivariado X toma valores en el espacio de características $\mathcal{F} = \mathbb{R}^d$. Sin embargo, en clasificación funcional $\mathcal{F}$ es un espacio de dimensión infinita, por ejemplo, $\mathcal{F} = L^2[a, b]$ (recordamos que las variables que pertenezcan a estos espacios de dimensión infinita se denotarán con letras caligráficas, por ejemplo χ).La mayoría de estos métodos de clasificación funcional (k-NN, k-means, SVM, kernel, etc.) son adaptaciones al espacio de dimensión infinita de metodologías "clásicas" en la clasificación multivariada, aunque también hay métodos desarrollados específicamente para el caso funcional. A continuación se profundiza un poco más en estos problemas y se comentan algunos de los métodos más utilizados en la actualidad partiendo de [5] y utilizando otros textos para completar puntos concretos. Se hará especial hincapié en la clasificación supervisada, objetivo principal de este trabajo.

2.3.3.1. Clasificación No Supervisada

En principio, la clasificación no supervisada o *clustering*, es un problema más complejo que la clasificación supervisada ya que trata de hacer grupos lo más homogéneos posibles, y a la vez lo más distintos entre sí, sin conocer ninguna información adicional sobre los datos, ni siquiera el número de grupos. Esto hace que ni siquiera el objetivo esté claro. En el caso discriminante parece obvio que se intentará reducir el error de clasificación, pero en el no supervisado puede que no haya un único criterio para definir la partición óptima. Para cubrir esta carencia de un objetivo general, Fraiman et al han propuesto una idea, y por consiguiente una guía para la elección del número de grupos [39]. Esta propuesta define tres cantidades en función de medidas de dispersión de la distribución de los elementos condicionadas a las distintas particiones. La primera mide la dispersión de los grupos en relación con la dispersión global, pero no es suficiente ya que se puede disminuir constantemente si se aumenta el número de grupos. Las otras dos tratan de compensar el efecto de añadir más grupos teniendo en cuenta las ganancias relativas, de modo que la mejor partición será aquella que minimice las tres cantidades.

Una vez seleccionado el número de grupos hace falta un método. En el caso multivariado hay multitud de ellos destacando dos enfoques sobre los demás: los métodos basados en *K-means* y los jerárquicos. Ambos enfoques han tenido su adaptación al entorno funcional.

K-means es un clásico en el ámbito del clustering. Muy estudiado y con multitud de variantes y adaptaciones, fue uno de los primeros métodos en adaptarse al caso funcional. Se pueden encontrar demostraciones de consistencia, por ejemplo, en [74]. La idea es sencilla, se trata de que los elementos estén en el grupo a cuya media tienen una menor distancia, por ejemplo, en un espacio normado y para un K fijo, se busca una asignación $f : \mathcal{F} \longrightarrow 1, \cdots, K$ que define la partición

de modo que $G_k = f^{-1}(k)$ con $k = 1, \cdots, K$, y los G_k son disjuntos entre sí y cubren $\mathcal{F}$. Sean $h_1, \cdots, h_K \in \mathcal{F}$ los centros de los grupos, el objetivo de K-means será minimizar la expresión

$$E\left(\min_{i=1,\cdots,K} \parallel \mathcal{X} - h_i \parallel^2\right)$$

o en su versión empírica

$$\frac{1}{n}\sum_{i=1}^{n} \min_{k=1,\cdots,K} \parallel \mathcal{X}_i - h_k \parallel^2 .$$

Una vez asignados los elementos se recalculan los centros hasta que se estabiliza. Hay distintas formas de estimar los parámetros y han surgido distintos estudios y variaciones, entre ellas cabe citar los resultados de Biau et al. sobre K-means en espacios de Hilbert [11], el ITKM (Impartial Trimmed K-Means) de Cuesta-Albertos y Fraiman para datos en espacios de Banach uniformemente convexos [21] o los *Puntos principales funcionales* propuestos por Simizu y Misuta [90]. Se puede encontrar un buen resumen de muchos de estos aspectos en [5].

Por su parte, los métodos jerárquicos, ya sean aglomerativos o disociativos, tienen por objetivo agrupar conjuntos o romperlos respectivamente, de forma que en los nuevos grupos se minimice una distancia o se maximice una medida de similitud. En principio estos métodos son perfectamente adaptables al caso funcional. Basta con definir una distancia o una medida de similitud adecuada. En esta línea, en la tercera parte de [35] se propone un método jerárquico disociativo desde una aproximación no-paramétrica. La metodología consiste en la partición iterativa en grupos cada vez menos heterogéneos, midiendo esta heterogeneidad en función de la cercanía de varias medidas de centralidad (media, mediana y moda).

2.3.3.2. Clasificación Supervisada

En primer lugar se abordará la definición formal de un problema de clasificación supervisada en dimensión infinita. Hay mucho puntos de coincidencia con el análisis discriminante multivariado, pero hay importantes diferencias que hacen necesaria la adaptación de los métodos, tanto formalmente como en los aspectos computacionales, ya que por lo visto anteriormente el problema en dimensión infinita es más complejo. Tras las definiciones y conceptos básicos se tratan algunos de los principales métodos de clasificación funcional, tanto por su importancia histórica como por su uso en la actualidad.

Sea $D_n = \{(\chi_i, Y_i), 1 \leq i \leq n\}$ el conjunto de entrenamiento con observaciones independientes que siguen la distribución de la variable aleatoria $\chi \in \mathcal{F}(\sim L^2[a,b])$ con $Y \in \{0,1\}$. Se define un clasificador g como la aplicación

$$g : \chi \longrightarrow \{0,1\}$$

construida para minimizar el error de clasificación $P(g(\chi_i) \neq Y_i)$ para todo i. La conocida como Regla de Bayes define el clasificador del mismo nombre del siguiente modo:

$$g^*(\chi) = I_{\{\eta(\chi)>1/2\}}$$

donde $\eta(x) = E(Y|X = x)$, es decir, la esperanza de que la observación pertenezca a una cierta clase dada la observación, e I es la función indicatriz. Este clasificador es óptimo en el sentido de que minimiza el error de clasificación definido anteriormente [26] y el objetivo de todos los clasificadores que se construyan será acercarse lo más posible al de Bayes. El problema de g^* es que sólo se puede definir en contadas ocasiones por la dificultad de dar una expresión para η.

Por tanto, los clasificadores que se construyen tendrán siempre como cota de error, el error Bayes. Estos clasificadores $g_n(x) = g_n(x; D_n)$ se construyen en la práctica a partir de la información del conjunto de entrenamiento. El error condicionado de estos clasificadores es $L_n = P(g_n(\chi) \neq Y|D_n)$ y debe tender al óptimo $L^* = P(g^*(\chi) \neq Y)$. De hecho, para una secuencia de clasificadores g_n se mide la consistencia en función de su convergencia al error Bayes. Así, si $L_n \longrightarrow L^*$ en probabilidad, o lo que es lo mismo, $E(L_n) \overset{n\to\infty}{\longrightarrow} L^*$, se dice que g_n es débilmente consistente. Del mismo modo, si converge casi seguramente en lugar de en probabilidad, se dice que la secuencia es fuertemente consistente. Para terminar, hay múltiples formas de crear estas aplicaciones, dos de las más comunes son las siguientes:

- Los llamados métodos de plug-in sustituyen η y aplican el clasificador de Bayes con esta nueva η. De nuevo, al desconocer la distribución conjunta de (χ, Y) no se puede calcular el error de g, teniendo que recurrir a un estimador llamado riesgo empírico y definido de siguiente modo:

  $$\widehat{L}_n = \widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{g(\chi_i)\neq Y_i\}}$$

- A partir de esta expresión se pueden definir los clasificadores basados en riesgo. La idea es elegir una clase C de clasificadores (con una estructura simple o cualquier otra propiedad interesante) y elegir como clasificador al que solucione el problema de minimización del riesgo empírico:

  $$g_n^* = argmin_{g\in C}\widehat{L}_n(g).$$

  Los clasificadores así construidos suelen dar mejores resultados que los de plug-in. Esto puede ser debido a que no tienen que estimar η, que ya es por sí solo un problema complicado.

Discriminante Lineal Este método es uno de aquellos en los que el caso multivariado difiere del funcional.

- *Multivariado*: Aunque este caso es ampliamente conocido, conviene recordar su planteamiento para compararlo con el caso funcional. Sea $\mathcal{F} \in \mathbb{R}^d$, P_0, P_1 poblaciones tales que $\Sigma_0 = \Sigma_1$ y $\mu_0 \neq \mu_1$ (sus matrices de covarianzas y sus medias respectivamente). La regla se basa en una transformación lineal β que hace $\mu_i \longrightarrow \beta^t \mu_i$ y se asigna el individuo x a la clase cuya media proyectada, $\beta^t\mu_0$, $\beta^t\mu_1$ sea más cercana a su propia proyección, βx. La proyección trata de maximizar la distancia entre las medias de las dos clases y minimizar la varianza dentro de cada clase sujeto a que $\beta^t\Sigma\beta = 1$, donde Σ es la varianza total. Así se elige la transformación:

  $$\beta^* = \underset{\beta \in \mathbb{R}^d}{argmax} = \frac{(\beta^t\mu_0 - \beta^t\mu_1)^2}{\beta^t\Sigma\beta}$$

  En este caso, la solución es equivalente a minimizar la distancia de Mahalanobis y se puede concluir que x se asigna a P_0 si

  $$(x-\mu_0)^t\Sigma^{-1}(x-\mu_0) < (x-\mu_1)^t\Sigma^{-1}(x-\mu_1)$$

- *Funcional*: Ahora $\mathcal{F} \in L^2[a,b]$ y denotamos el producto escalar en $L^2[a,b]$ de la forma $\langle f, g\rangle = \int_a^b fgdt$. Sea $x \in \mathcal{F}$, el clasificador se obtendrá proyectando el elemento x de dimensión infinita en $\mathbb{R}$ y comparándolo con las proyecciones de las medias, ahora $\mu_j = E(\chi|Y = j)$ y sus proyecciones $\langle \beta, \mu_j\rangle$. El objetivo vuelve a ser maximizar las distancias entre las medias, pero ahora la restricción es $\int_a^b \beta(t)\langle\beta, V(t,.)\rangle dt = 1$, con $V(s,t) = Cov(\chi(t), \chi(s)|Y = j)$. Siguiendo la misma filosofía se elige la proyección

  $$\beta^* = \underset{\beta \in \mathcal{F}}{argmax} = \frac{Var(E(\langle\beta,\chi\rangle|Y))}{E(Var(\langle\beta,\chi\rangle|Y))}$$

  El problema en dimensión infinita, es que no hay solución, ya que $V(s,t)$ no es en general invertible, por lo que hay que recurrir a aproximaciones. Una solución es la discretización de χ de forma que la integral se transforma en un sumatorio. Sin embargo, valores consecutivos de la discretización están muy correlados y generan una matriz de covarianzas casi singular y por tanto, muy difícil de resolver numéricamente (esto se podría solucionar añadiendo una matriz de penalización diagonal [108]). Recientemente Li y Yu han desarrollado en este sentido un análisis discriminante lineal sobre segmentos de curvas aplicando técnicas SVM [61]. La otra forma de encontrar una solución en la práctica es recurriendo a una base de funciones, donde el problema, como se ha dicho anteriormente, es elegir la base adecuada.

k-NN El método de los vecinos más próximos es válido tanto para datos funcionales como multiatributo. Sea $\mathcal{F}$ un espacio métrico, se clasifica χ en función de los k ejemplos de entrenamiento

más próximos según la métrica de $\mathcal{F}$. Los empates se pueden decidir aleatoriamente o con algún otro procedimiento, y k juega en este método el papel de parámetro de suavizado. Se puede considerar como un método de plug-in tomando

$$\eta(\chi) = \eta_n(\chi) = \frac{1}{k}\sum_{i=1}^{n} I_{\{\chi_i \in k(\chi)\}} Y_i$$

y entonces

$$g_n(\chi) = I\{\eta_n(\chi) > 1/2\}$$

donde $\chi_i \in k(\chi)$ hace referencia a si la i-ésima observación pertenece a los k vecinos más próximos a χ. La selección del parámetro k se suele hacer minimizando el error por validación cruzada. La única diferencia entre funcional y multivariado es la elección de una distancia de forma adecuada, la cual viene definida por el espacio funcional y usualmente se toman $L^2[a,b]$ con la métrica usual o $C[a,b]$ con la métrica del supremo, si bien otras aproximaciones pueden ser interesantes.

k-NN es un método ampliamente utilizado, en [5] se comentan algunas de las propiedades de consistencia del método, aunque para una mayor profundización se puede consultar [3] donde basándose en las trabajos de Cérou y Guyader (2006) se prueba la consistencia del algoritmo para clasificación de datos funcionales con ciertas distribuciones. Además, en este artículo y en [4] se analiza el comportamiento de k-NN en relación a otros clasificadores en distintos problemas, y aunque no es óptimo en todos los casos, se propone como método de referencia por su buen comportamiento global. Ya sea por estos buenos resultados globales, sus buenas propiedades matemáticas o su sencillez, k-NN es un método muy popular y utilizado, y sobre el que se sigue investigando como muestran los trabajos de Zhang y Zhou, y Younes, Abdallah y Denoeux, con versiones para clasificación multietiqueta [111, 107], o con métodos híbridos como en el caso de [2].

Kernel En el sistema anterior se fijaba el número de elementos a considerar. En los métodos de Kernel se fija una distancia h y "votan" todos los elementos que no disten más de ésta. Una vez definida la métrica con una distancia d, el individuo χ pertenecerá a la población P_0 si

$$\sum_{i=1}^{n} I_{\{Y_i=0, d(\chi_i,\chi)\leq h\}} > \sum_{i=1}^{n} I_{\{Y_i=1, d(\chi_i,\chi)\leq h\}}$$

De forma más general, si definimos una función de Kernel $K : [0,\infty) \longrightarrow [0,\infty)$ decreciente, por ejemplo, $K(x) = e^{-x^2}$ el núcleo gaussiano, o $K(x) = (1-x^2)I_{[0,1]}$ el núcleo de Epanechnikov, $g_n(\chi)$ asigna χ a la clase 0 si

$$\sum_{i=1}^{n} I_{\{Y_i=0\}} K\left(\frac{d(\chi_i,\chi)}{h}\right) > \sum_{i=1}^{n} I_{\{Y_i=1\}} K\left(\frac{d(\chi_i,\chi)}{h}\right)$$

Los métodos de Kernel pueden verse de nuevo como un método de plug-in definiendo:

$$\eta_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(\chi_i,\chi)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{d(\chi_i,\chi)}{h}\right)}$$

Las dificultades de estos métodos radican en la elección de la distancia de un modo similar a k-NN (a veces es mejor utilizar semi-métricas [$d(x,x) = 0$] basadas en PCA o PLS para funciones poco suaves (2.2.2, página 16)) y en la selección de h. Para ajustar este parámetro se puede realizar una validación cruzada que disminuya el error como en el caso anterior, sin embargo, la complicación es mayor al ser h un parámetro continuo. En la monografía de Ferraty y Vieu [35] se realiza un profundo estudio de los métodos de Kernel y abordan este problema. Para el caso de dimensión finita se ha probado que esta regla de clasificación es universal fuertemente consistente si la función de kernel es regular y se cumple que cuando $n \to \infty$, $h \longrightarrow 0$ y $nh^d \longrightarrow \infty$, donde d es la dimensión del problema. Sin embargo, el estudio de las condiciones suficientes de consistencia en el caso funcional es un problema en estudio, recientemente Abraham et al. han ofrecido algunos resultados en cuanto a la consistencia débil [1].

PLS Los mínimos cuadrados parciales (PLS) son una técnica de reducción de dimensiones que ha ganado mucha presencia recientemente en el ámbito multivariado. La idea que subyace es similar a la de PCA, proyecciones en las direcciones de máxima variabilidad, con la diferencia de que PLS tiene en cuenta la respuesta Y, lo que lo hace mejor que este para la clasificación [6]. En el caso funcional, este método cada vez aparece en más referencias con datos funcionales, y aún teniendo su origen en la regresión se aplica cada vez más en clasificación [35, 4]. Especial mención merece el trabajo de Preda et al. [77] en el que se aplica PLS sobre datos funcionales en combinación con el discriminante lineal. Sin embargo, pese a pequeños detalles sobre la formulación teórica, en la práctica PLS se aplica sobre una discretización del dato funcional como si fuera multivariado, por ello hablaremos más extensamente del método en la sección dedicada a ello (3.1, página 38).

Reproducing Kernel La teoría de kernel reproductores surge del análisis matemático para pasar después a la estadística. Ésta es una teoría de transformación que asocia una función kernel definida positiva con un espacio de Hilbert. Como otras teorías, los kernel reproductores se fundamentan en que problemas complejos en un cierto espacio pueden ser fácilmente resolubles en otro, y la solución óptima en éste suele serlo también en el primero [9]. Estos sistemas empezaron a cobrar importancia con los trabajos de Parzen en los años 50, dejando de ser una herramienta oscura, y se han ido estudiando y perfeccionando hasta irrumpir con fuerza en el mundo de la clasificación de patrones de la mano de los SVM de Vapnik y otros trabajos.

La filosofía de estos métodos es simple, la idea es utilizar una función kernel para llevar un conjunto de observaciones a un espacio de Hilbert donde la distancia entre las observaciones venga determinada por el kernel. Asignando esta distancia se pueden utilizar las herramientas tradicionales.

Centrándonos en el problema de clasificación y formalizando estas ideas se define $K : \mathcal{F}\text{x}\mathcal{F} \longrightarrow \mathbb{R}$ definida positiva como un *Kernel Reproductor* de un espacio de Hilbert (H,$\langle ., . \rangle$) de funciones reales definidas en $\mathcal{F}$ que cumple:

1. Para todo $x \in \mathcal{F}$, $K(., x) \in H$.
2. Para todo $x \in \mathcal{F}, \varphi \in H$, $\langle \varphi, K(.x) \rangle = \varphi(x)$.

Los espacios de Hilbert en los que se da esto reciben el nombre de RKHS (Reproducing Kernel Hilbert Space) y las condiciones para que un espacio de Hilbert sea RKHS se explican en [9]. Estas propiedades hacen que podamos operar con elementos en espacios de Hilbert mediante productos escalares en los espacios de partida y así, como se ha comentado antes, utilizar el Kernel $K(x_i, x_j)$ para meter los puntos $x_i, \ldots, x_n$ en H con $K(x_i, x_j)$ como distancia entre x_i, x_j y utilizar otros métodos.

Como clasificadores se estima una $\hat{\eta} \in H_K$ (RKHS asociada al kernel K) minimizando el riesgo empírico regularizado

$$\frac{1}{n} \sum_{i=1}^{n} C(\chi_i, Y_i, \hat{\eta}(\chi_i)) + J(\hat{\eta})$$

donde C es una función de coste convexa respecto al tercer argumento y $J(\hat{\eta})$ es un término de penalización. Una elección típica es tomar $C(\chi_i, Y_i, \hat{\eta}(\chi_i)) = (Y_i - \hat{\eta}(\chi_i))^2$ y $J(\hat{\eta}) = \lambda \parallel \hat{\eta} \parallel_K^2$, para un cierto parámetro de regularización $\lambda > 0$. En nuestro caso también puede definirse la función de coste como $C(\chi_i, Y_i, \hat{\eta}(\chi_i)) = -Y \log \hat{\eta}(\chi_i) + \log(\frac{1}{1-\hat{\eta}(\chi_i)}$. Con esta función, la solución al problema de minimización es:

$$\hat{\eta}(x) = \sum_{i=1}^{n} \alpha_i K(x, \chi_i),$$

con los coeficientes reales. Una elección estándar para K es el núcleo gaussiano $K(x, y) = e^{\frac{1}{\sigma_k^2}\|x-y\|^2}$, con la distancia euclídea. Aunque estos métodos son prometedores todavía se están estudiando sus efectos y aplicaciones a la clasificación supervisada.

SVM RKHS es más una metodología general que un método definido. Incluso elegidos el kernel y C aún hay que ajustar distintos parámetros. El más conocido de los procedimientos que aplica la filosofía de RKHS son las máquinas de vectores soporte (SVM). Los SVM fueron introducidos por

Vapnik que fijo el marco teórico en el contexto multivariado [95]. Estos primeros trabajos se centran en el caso binario, que sigue siendo el más representativo aunque han proliferado los trabajos, generalizaciones y modificaciones del método hasta convertirse en referente. En el caso binario, donde las observaciones se denotan por (x_i, y_i) con $x_i \in \mathbb{R}^d$ e $y_i \in \{-1, 1\}$, se trata de obtener un clasificador de la forma $\phi_n(x) = \text{Signo}\{\langle w, \psi(x)\rangle_H + b\}$, donde $\psi : \mathbb{R}^d \longrightarrow H$ es una extensión a un espacio de Hilbert. La idea es que en el espacio de Hilbert el problema se separará más fácilmente. Para ajustar los coeficientes w y b se resuelve el problema de maximización del margen entre las observaciones de las distintas clases y la frontera de decisión. En un primer momento no se permitían muestras mal clasificadas, pero esto era muy poco flexible y no se ajustaba a la práctica, por lo que después se permitieron, quedando el problema de maximización del siguiente modo:

$$\min_{w,b,\epsilon} \parallel w \parallel_H^2 + C \sum_{i=1}^{n} \epsilon_i$$

tal que, $y_i(\langle w, \psi(x)\rangle_H + b) \geq 1 - \epsilon_i$ y $\epsilon_i \geq 0$, $1 \leq i \leq n$. Este es un problema complicado, pero tiene una formulación dual en función de los productos internos $\langle \psi(x_i), \psi(x_j)\rangle_H$ que mediante la utilización de los kernel reproductores hace innecesario conocer el propio espacio H y el producto interno de forma explícita, ya que elegido el kernel como en la sección anterior, existe un H y un ψ con las características comentadas tales que $\langle \psi(x_i), \psi(x_j)\rangle_H = K(x_i, x_j)$. En este "truco" radica la potencia del método y es la base de su éxito, además ya hay resultados de consistencia universal bajo ciertas hipótesis.

La extensión funcional surge de forma natural. Todo el mecanismo se puede adaptar a datos en un espacio funcional $\mathcal{F}$, en lugar de $\mathbb{R}^d$, pasando a las versiones continuas (sumatorio por integral, por ejemplo), sin embargo, la naturaleza de los datos funcionales hacen que no sea aplicable en muchos casos esta adaptación trivial (o no consiga buenos resultados) e invalida los resultados de consistencia. En primer lugar hay que tener en cuenta el preprocesado de datos mediante discretizaciones, proyecciones a una base finita de funciones o cálculo de derivadas. En cuanto a los kernel, la mayoría de los clásicos se adaptan al caso funcional, como el gaussiano $K(f, g) = e^{-\sigma\|f-g\|^2}$, o el polinómico $K(f, g) = (1 + \langle f, g\rangle)^D$, sin embargo, la naturaleza funcional de los datos puede utilizarse para definir nuevos kernels, por ejemplo combinando los clásicos con una función de mapeo. A nivel práctico, los datos nunca son funciones perfectas sino un vector de los valores en una partición, por lo que a menudo se toma la solución simple de aplicar el modelo de SVM estándar. Discusiones sobre estas ideas, ejemplos de buenas adaptaciones, resultados de consistencia en el caso funcional y experimentos en espectrometría y reconocimiento de voz se muestran en los artículos de Rossi y Villa [87, 99] entre otros.

Medidas de Profundidad En el punto 2.2.3, página 18, se han visto las dificultades para medir la profundidad en datos funcionales y algunas definiciones, ahora comentaremos brevemente cómo

utilizar estas medidas para clasificar, pueden encontrarse referencias en [5] y algunos resultados en [24] La idea subyacente es que si podemos medir la profundidad de un dato en la nube de puntos, podremos definir la mediana del conjunto como el dato más profundo y utilizar esta información para clasificar. Este es un paradigma general dentro de la clasificación ya que tenderemos a clasificar un nuevo dato en la clase entre cuyos representantes el nuevo elemento sea más profundo. Sea F la función de distribución de probabilidad de la población en la que se encuentra el individuo, en la recta real se puede definir la profundidad de un dato de manera sencilla como

$$D(x) = F(x)(1 - F(x))$$

o bien,

$$D(x) = \text{mín}(F(x), 1 - F(x))$$

y podemos reducir el problema a ese caso. En general, si tenemos una medida $D(P_i, x)$ de la profundidad de x en P_i se puede utilizar en clasificación definiendo como clasificador

$$g(x) = I_{\{D_{n1}(x) > D_{n0}(x)\}}$$

siendo $D_{ni}(x)$ la estimación empírica de $D(P_i, x)$ obtenida del conjunto de entrenamiento. Sobre estas ideas surgen nuevas aproximaciones que aplican la noción de profundidad en datos funcionales como una herramienta en problemas de clasificación. En esta línea se encuentran los trabajos de Cuevas y Fraiman [25] proponen tomar una profundidad media en espacios de Banach reales y separables, o los de López-Pintado y Romo [66] que plantean una aproximación gráfica para la profundidad.

Otras Consideraciones Los métodos aquí comentados son, en función de lo estudiado, los más representativos por su grado de utilización, su desarrollo a lo largo del tiempo, las mejoras o variaciones surgidas de ellos, o la profundidad de los estudios realizados. Hay muchos más: desde adaptaciones recientes de métodos consagrados en otros campos como las redes neuronales [36, 86], a soluciones concretas en problemas específicos como la aproximación por segmentación en [61] o transformaciones afines para curvas en dos y tres dimensiones en el tratamiento de imagen [33], pasando por metodologías para la preparación de los datos antes del propio proceso discriminante como en [8], o ampliaciones y variaciones de los métodos ya citados, por ejemplo en [35] hay bastante información sobre algunos de los métodos no-paramétricos viéndolo como un problema de regresión con las clases como respuesta. Sin embargo, no han sido incluidos por redundancia, claridad o por no ser lo suficientemente representativos o estar poco estudiados.

Capítulo 3

El problema de la dimensión

Como ya se ha comentado, las funciones acaban manejándose en la práctica en una versión discretizada de forma que, aunque la discretización sea una mera aproximación, al final tenemos un vector de datos en $\mathbb{R}^D$ en el que se podrían aplicar técnicas de reducción de dimensión de análisis multivariado. El interrogante es si métodos no pensados originalmente para funciones son aplicables o pueden dar buenos resultados. No está claro que haya una respuesta global y que todos los enfoques sean trasladables al caso funcional, pero es obvio que existe una estrecha relación y se podría pensar en la utilización o adaptación de estos métodos. Prueba de ello es la utilización en el ámbito funcional de PLS [77] o la redefinición de PCA en FPCA (2.3.1, página 22). Sin embargo hay multitud de enfoques que no se han aplicado en FDA. En particular, los métodos de selección de variables, por sus características y representatividad, parecen buenos candidatos a "dar el salto" a las funciones, ya sea en una versión "pura" (se trabaja con discretizaciones) o mediante alguna posible adaptación. Al no haber apenas bibliografía en este sentido, en próximas secciones de este trabajo se motiva la utilización de un algoritmo de selección de variables estándar sobre datos funcionales y se estudian resultados preliminares en el marco de la clasificación supervisada. Para ello, en este capítulo se hace un breve repaso de las técnicas de reducción de dimensión multivariadas intentando dar una visión general del panorama actual pero sin entrar en detalles que se alejan de los objetivos del presente documento. Se comentarán con mayor profundidad los dos algoritmos elegidos para realizar las pruebas: mRMR por parte de la selección de variables y PLS como método de proyección para hacer una pequeña comparativa.

La *maldición de la dimensionalidad* es una expresión acuñada por Bellman en los años 60 para referirse al hecho de que el número de patrones necesarios para estimar una cierta función con un determinado grado de precisión crece exponencialmente con el número de variables [12]. Actualmente, son muchos los problemas que presentan un altísima dimensionalidad con un reducido número de muestras, y están sujetos, por tanto, a la maldición de la dimensionalidad: datos fun-

cionales, microarrays o análisis de texto entre otros. Otras veces será útil o necesario aumentar artificialmente la dimensión de los datos para conseguir una mejor separabilidad. En ocasiones el problema estará gobernado por unas pocas variables latentes que lo simplifican. En todos estos casos se pone de manifiesto la necesidad de la reducción de dimensiones por distintos motivos: disminuye el error de clasificación al quedarse con la información representativa, permite abordar problemas de mayor dimensionalidad con menor coste, pone de manifiesto estructuras y relaciones latentes o permite modelos más sencillos y comprensibles. El objetivo, y a la vez el problema, es reducir la dimensión de los datos minimizando la pérdida de información útil. Para solventar este problema, el análisis multivariado ofrece dos vías, la extracción de características y la selección de variables. El primer conjunto de métodos trata de reducir la dimensión del problema mediante la proyección (ya sea lineal o no-lineal) de los vectores D-dimensionales a un espacio de dimensión d con $d \ll D$. Por su parte, los métodos de selección de variables reducen la dimensión quedándose con un subconjunto de las variables originales del problema.

3.1. Extracción de características

Se tiene el conjunto de datos $\{x_n\}_{i=1}^N$ con $x_i \in \mathbb{R}^D$ siguiendo una determinada distribución de probabilidad. La idea básica que subyace en estos métodos es que el conjunto de datos puede ser representado o descrito mediante una muestra que reside en una variedad de dimensión menor. El objetivo de la reducción de dimensiones es encontrar la expresión en coordenadas de esa variedad para proyectar los datos en ella. Por ejemplo, PLS obtiene variedades lineales (subespacios vectoriales) generados por las primeras componentes del método. Con estas consideraciones se concluye que todo problema tiene una dimensión intrínseca que se define como el número de variables independientes que explican satisfactoriamente el problema. La dimensión intrínseca d, será la dimensión de la variedad que recoja los datos de una cierta distribución desconocida en el espacio D-dimensional $(D > d)$ satisfaciendo ciertas restricciones [17]. La formalización de estas ideas dista mucho de ser trivial e involucra diferentes nociones de dimensión y complejidad en las que aparece una sugestiva interacción entre la estadística y la computación.

Hay una gran variedad de métodos de reducción de dimensiones de este tipo [94]. Se podría hablar de ellos en función de su campo de aplicación, por ejemplo en Análisis Semántico Latente (LSA) dedicado a la extracción de características en textos mediante la Descomposición en Valores Singulares (SVD) de matrices. También se podrían explicar según su objetivo y hablar de los que maximizan la covarianza con la clase como PLS o los que mantienen las distancias como Diffusion Maps. Se podrían agrupar por "familias" o derivaciones, como es el caso de Kernel PCA que reformula el análisis de componentes principales calculando los autovectores de la matriz de kernel

en lugar de los de la de correlaciones ($K(x_i, x_j)$ en vez de $Cov(x_i, x_j)$). También sería posible presentarlos por contraste, de este modo, el Análisis de Componentes Independientes (ICA) podría contrastarse con PCA como alternativa no-lineal que observa la independencia estadística mediante la utilización de la información mutua en lugar de la correlación lineal de PCA. En resumen, se podrían clasificar o estudiar de muy distintas formas, como se puede apreciar en la literatura, pero sobre todas estas jerarquías sobresalen dos divisiones, métodos locales frente a globales, y sobre todo, lineales frente a no-lineales. Los primeros, cuyo máximo representante es PCA, asumen que la estructura de los datos está en un subespacio lineal, mientras que los segundos no toman esta hipótesis y tratan de sacar provecho de las relaciones no lineales que se dan en la práctica. A su vez, estos métodos pueden subdividirse entre globales y locales. Los sistemas globales tratan de preservar características globales de los datos en el espacio de dimensión menor, como el Escalado Multidimendional (MDS) o Isomap que tratan de conservar las distancias entre los pares de puntos (la distancia euclídea y la medida por las geodésicas de la variedad respectivamente). Por su parte, los métodos locales intentan preservar las propiedades locales arguyendo que éstas determinan las globales. Este es el caso de LLE (Local Linear Embedding), similar a Isomap, pero observando subgrafos locales en lugar del grafo global.

Todos los métodos tienen sus ventajas y sus inconvenientes y no hay ninguno claramente superior a los otros. Por ejemplo, los sistemas locales son más vulnerables a la maldición de la dimensionalidad cuando la dimensión intrínseca es grande (caso frecuente), son incapaces de detectar discontinuidades en la variedad y pueden sobreajustar (sobre todo en presencia de ruido). Sin embargo, algunas de estas debilidades también las pueden presentar algoritmos globales que construyan grafos sobre la variedad. Los métodos basados en kernel no sufren estas desventajas pero una elección inadecuada de la función de kernel impedirá la correcta modelización de la variedad.

El estudio de todos estos métodos, aunque interesante, sobrepasa el alcance y los objetivos de este trabajo por lo que se abordará sólo parcialmente a través de uno de los sistemas. Aunque parece que las soluciones no-lineales deberían ser superiores al tener en cuenta la no-linealidad de los datos, en la literatura sólo mejoran claramente a los modelos lineales en conjuntos de datos sintéticos, mientras en con datos reales mejoran mínimamente o son superados. Una muestra de ello se encuentra en [94] donde se comparan 12 métodos no-lineales con PCA resultando éste último vencedor en 3 de los 5 conjuntos de datos reales que se estudian. Por esto, y por la sencillez de los modelos lineales se ha elegido un método de la familia de Projection Pursuit (PP), en concreto PLS. Puede encontrarse información actual sobre distintos métodos en [62].

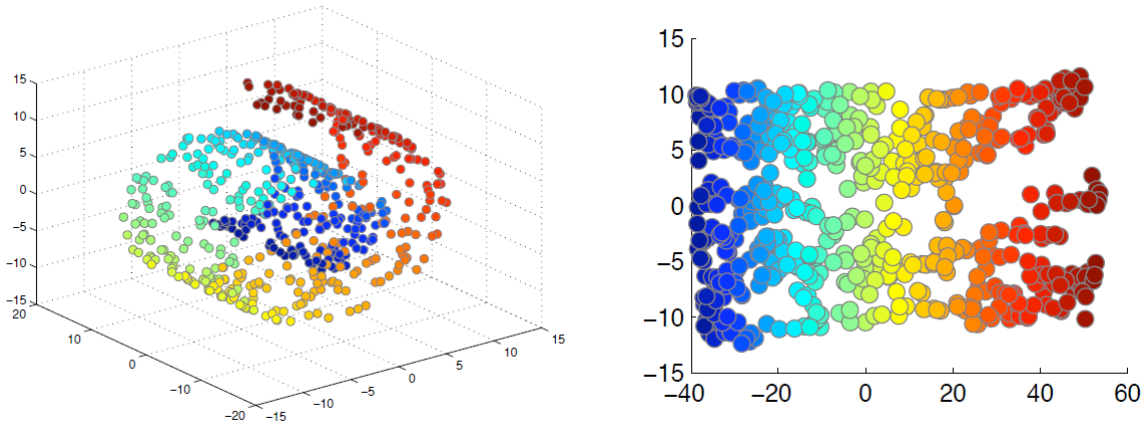

Figura 3.1.1: La imagen de la izquierda muestra el conjunto sintético conocido como *swiss roll*. Es un típico ejemplo en el que los métodos no lineales se imponen a los lineales ya que LLE o Isomap (imagen derecha) lo resuelven perfectamente mientras que un método lineal como PCA es incapaz. Sin embargo, otros métodos no-lineales como MDS no pueden resolverlo, mostrando que no hay un método ni un paradigma óptimo en todos los casos.

Los métodos PP buscan la mejor proyección a una dimensión muy baja. Muchos modelos, incluyendo PCA y PLS, pueden verse como casos particulares de éstos. PP es una técnica no supervisada que elige las proyecciones lineales y ortogonales de un espacio de alta dimensión que optimizan una cierta función objetivo llamada índice de proyección (en PCA la varianza y en PLS la covarianza con la clase), tratando de explotar la habilidad de las personas para descubrir patrones y estructuras en dimensiones bajas. Los distintos criterios para definir la proyección óptima dan origen a la gran variedad de métodos con esta filosofía, pero el fin último en todos ellos es encontrar la estructura subyacente que más facilite las operaciones posteriores (clasificación, modelización, selección, o cualquier otra cosa).

En general, el índice de proyección Q es una función real del espacio de distribuciones en $\mathbb{R}$

$$Q : f \in L^2(\mathbb{R}^d) \longrightarrow q = Q(f) \in \mathbb{R}$$

Donde $f = F_A$, la distribución de la proyección (con la matriz A, de dimensión D x K) de una variable aleatoria D-dimensional X con distribución original F. Esta proyección de X a $\mathbb{R}^d$ es una nueva variable aleatoria $Y = A^t X$, y es en este nuevo espacio de llegada en el que queremos maximizar o minimizar algún criterio recogido por el funcional Q. PP trata de encontrar las direcciones de proyección a_i, que producen un óptimo de Q dada la distribución F. Además, para hacer el problema de optimización independiente del tamaño de los vectores de proyección y obtener direcciones incorreladas las direcciones de proyección han de ser ortonormales, es decir, $a_i^t a_j = \delta_{ij}$

Los índices de proyección fueran deben seguir ciertas reglas como ser invariantes ante traslaciones, cambios de escala y transformaciones afines o diferenciables para poder hacer descenso por gradiente. Algunos de estos índices son la media, la varianza, la información de Fisher, la entropía de Shannon, funciones basadas en polinomios de Hermite o de Legendre,... [17]

Entre los distintos métodos que tienen cabida en la familia PP, destacan por su difusión PLS y sobre todo PCA. El Análisis de Componentes Principales es muy popular y se ha escrito mucho sobre él. Además de ser bien conocido es relativamente sencillo y su interpretación es más fácil que la de PLS. Sin embargo, hemos elegido PLS por el sesgo del presente trabajo hacia la clasificación supervisada y por las características de los datos funcionales ya que PLS ha probado desenvolverse con soltura en muestras pequeñas con muchas variables [15].

PLS

PLS es un método originario de la química, en concreto de la quimiometría: parte de la química que aplica métodos matemáticos y estadísticos sobre datos químicos. Se pueden encontrar abundantes referencias en publicaciones como *Journal of Chemometrics*. Fue planteado por primera vez por Wold en 1975 como método de regresión mediante el algoritmo iterativo NIPALS. Los orígenes del método son empíricos y surgió, como muchos otros, sin un fuerte fundamento teórico detrás. Sin embargo, los buenos resultados obtenidos hasta el momento han motivado su estudio formal, en esta línea se pueden consultar trabajos como [73] y [65]. En [41] se presenta el modelo en este marco junto con otros algoritmos de regresión. Los detalles del algoritmo pueden encontrarse en distintos artículos como el mismo [41] o [70]. En este sentido cabe destacar el trabajo de Vega Vilca [98], en el que se reconvierte el modelo orientándolo a clasificación. Como ya hemos dicho, el objetivo es Maximizar la covarianza con la clase, definida de la forma estándar:

$$Cov^2(X,Y) = E[(X - E(X))(Y - E(Y))]^2 = (\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}))^2$$

Según este planteamiento se trata al vector Y, por ejemplo de ceros y unos (en el caso general podrá ser una matriz) como una variable de regresión y se implementa el algoritmo NIPALS. De este modo se buscarán las K componentes que minimicen el error de:

$$\widehat{Y} = \beta_0 + \beta_1 T_1 + ... + \beta_K T_K$$

Ahora el objetivo es la clasificación, por lo que el algoritmo no devolverá unos modelos de regresión, sino un conjunto de componentes (proyecciones), T y una matriz de proyección Z para tratar los nuevos datos. El algoritmo para el caso de salidas unidimensionales fue concluido por Trygg [93] modificando la idea de Wold:

1. Inicialización: $Y(0) = Y$ estandarizado y $X(0) = X$ estandarizado.
2. Para k de 1 a d:
3. $w(k) = Cov[Y(k-1), X(k-1)]$
4. $w(k) = \frac{w(k)}{\|w(k)\|}$
5. $T_k = X(k-1)w(k)$
6. $v(k) = (\frac{T_k^t Y(k-1)}{T_k^t T_k})$; $b(k) = (\frac{T_k^t X(k-1)}{T_k^t T_k})$
7. $Y(k) = Y(k-1) - T_k v(k)$; $X(k) = X(k-1) - T_k b(k)$
8. Si $k = 1$ $z_1 = w(1)$ y si no $z_k = [Id - \sum_{j=1}^{k-1} z_j b(j)]w(k)$

Los índices entre paréntesis hacen referencia a la iteración mientras que los subíndices señalan las coordenadas. Este algoritmo extrae en cada iteración una nueva componente ortonormal a las anteriores de forma que se maximiza la función objetivo eligiendo como vector de proyección la covarianza entre los datos de entrada y la clase. Al final de cada iteración se actualizan las matrices X y Y eliminando la "información" obtenida con la nueva componente generada, esto es, se calculan los coeficientes de regresión asociados a cada una de ellas y se restan. Según este planteamiento, las componentes T_i son las proyecciones de los datos. La matriz Z es la matriz de proyección para nuevos datos. En la tesis [98] se detalla este sistema y se justifican las buenas propiedades de las componentes así extraídas. Además se propone una extensión multivariada y se estudian variaciones del algoritmo.

PLS adaptado a un sistema de reducción de dimensiones orientado a clasificación irrumpe con fuerza en el ámbito de la bioinformática para tratar los problemas de microarrays [15][38]. Una formulación del modelo como paso previo a la clasificación se encuentra en [6] donde PLS se muestra claramente superior a PCA. Está superioridad inicialmente "intuitiva" [6], ha sido demostrada formalmente en el caso binario [65] y comprobada empíricamente en distintos trabajos como [70]. Al igual que PCA, PLS puede verse como un método PP, donde $Q(W^t X) = Cov^2(Y, T)$ siendo Y el vector de salidas (clases) asociado a X y $T = W^t X$ la matriz de componentes (variables sintéticas). PLS intentará maximizar la covarianza de las proyecciones de los datos con la clase a la que pertenecen. Aquí encontramos la diferencia fundamental con PCA ya que PLS tiene en cuenta la información que da la clase, mientras que PCA contempla únicamente la varianza global de los datos.

En general, las salidas tendrán dimensión M, si bien, para hacer el estudio de clasificación se fija $M = 1$ (problema binario) para después generalizar. De esta forma, $T_{NxD} = (T_1, \cdots, T_D)$ tiene por columnas las componentes extraídas por PLS, que son ortonormales. De todas formas, considerando la matriz completa T no ganamos nada, ya que se vuelve a un espacio de dimensión D. Por este motivo se cogen sólo las primeras K componentes, $K \lll D$.

Como en la mayoría de los modelos de este tipo, la elección de K es uno de los mayores problemas y no existe una solución general. Este parámetro suele ajustarse en la fase de validación por sistemas estándar como la validación cruzada.

En el caso de PLS el problema es interesante porque permite ser abordado tanto por un algoritmo iterativo que va extrayendo las componentes una a una en función de las anteriores, como un problema de valores propios:

$$(a_{k+1}, b_{k+1}) = \underset{a \in \mathbb{R}^D, b \in \mathbb{R}^M, a^t A = 0}{argmax} \frac{Cov(a^t X, b^t Y)^2}{(a^t a)(b^t b)}$$

Donde a_k es la k-ésima componente correspondiente al k-ésimo autovalor, y A es la matriz de las $k-1$ componentes ya elegidas. Las matrices a diagonalizar serían:

$$H = \Sigma_{XY}\Sigma_{YX} = \begin{pmatrix} Cov(X_1,Y)^2 & Cov(X_1,Y)Cov(X_2,Y) & \cdots & Cov(X_1,Y)Cov(X_D,Y) \\ Cov(X_2,Y)Cov(X_1,Y) & Cov(X_2,Y)^2 & \cdots & Cov(X_2,Y)Cov(X_D,Y) \\ & \cdots & & \\ Cov(X_D,Y)Cov(X_1,Y) & Cov(X_D,Y)Cov(X_2,Y) & \cdots & Cov(X_D,Y)^2 \end{pmatrix}$$

en el caso de $M = 1$, y en general se tendrá en la diagonal la suma de las covarianzas al cuadrado de cada atributo con todas las salidas ya que:

$$\Sigma_{XY} = \begin{pmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) & \cdots & Cov(X_1,Y_M) \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) & \cdots & Cov(X_2,Y_M) \\ & \cdots & & \\ Cov(X_D,Y_1) & Cov(X_D,Y_2) & \cdots & Cov(X_D,Y_M) \end{pmatrix}$$

y $\Sigma_{YX} = \Sigma_{XY}^t$ su traspuesta. En este caso, las componentes se sacarían resolviendo el problema de autovalores hasta la dimensión K que queramos.

Sobre el modelo de PLS para clasificación han surgido variantes para adaptarlo a nuevas áreas, como a la clasificación de textos [110] o clasificación funcional [77], como para obtener nuevos enfoques, como Kernel-PLS [85] o los Oriented PLS propuestos por Liu y Rayens [65]. En ese mismo artículo se pone de manifiesto la estrecha relación entre los métodos de discriminante lineal

(LDA o Fisher) con la formulación de PLS como problema de autovalores. Pero los contrastes más interesantes y documentados en la literatura son, como ya se ha dicho, con PCA, que hacen que PLS sea preferible cuando el objetivo es la clasificación supervisada.

Por último, hay que comentar que las limitaciones de PLS en cuanto al número de componentes. Con el enfoque de valores propios, PLS tiene el mismo "problema" que Fisher, y es que si hay C clases sólo se pueden obtener C-1 autovalores distintos de cero [77] debido al rango de las matrices Σ_{YX} y Σ_{XY}, esto es obviado por algunos autores como en [65] pero debe tenerse en cuenta. Si se requieren más proyecciones se puede utilizar la versión iterativa. En este caso el límite será el rango de X (igual que en PCA) [70].

3.2. Selección de variables

La selección de variables aborda el problema de la reducción de la dimensión eligiendo el subconjunto de variables originales más informativo para mejorar el almacenamiento y procesamiento de los datos. El interés de estos métodos es que no alteran la representación original de los atributos preservando la semántica de las variables y, de este modo, facilitan la interpretación de los modelos por parte de expertos. La selección de variables puede aplicarse tanto en aprendizaje supervisado como no-supervisado. En este trabajo no centraremos en el caso supervisado (en particular, en la clasificación) pero se pueden encontrar información y referencias del no-supervisado en [30, 97, 64] e incluso del semi-supervisado (unas pocas muestras etiquetadas entre muchas sin etiquetar) en [63].

La selección de variables es un campo en continuo crecimiento debido a la irrupción de datos extremadamente grandes en distintas áreas. El objetivo de este trabajo no es hacer una revisión bibliográfica extensa de la selección de variables sino presentar la técnica de Mínima Redundancia y Máxima Relevancia (mRMR) y situarla dentro de la multitud de metodologías existentes. No obstante, para hacer esto será necesario motivar la utilización de la selección de variables, sintetizar los enfoques y metodologías más importantes, perfilar las tendencias y aportar referencias para una mayor profundización si fuera oportuno. La literatura es muy abundante pero a modo de guía sobre los conceptos básicos puede consultarse el trabajo de Guyon y Elisseeff [42] que combina distintas contribuciones para explicar, entre otros, los conceptos de relevancia, redundancia, las relaciones entre variables, métodos de validación, ranking de variables o búsquedas. Por su parte, Saeys, Inza y Larrañaga hacen una revisión muy completa de métodos de selección de variables en el contexto supervisado en [89]. En ambos artículos pueden encontrarse abundantes referencias para profundizar más. Para hacerse una idea de nuevas tendencias como la clasificación de información Web,

de textos, o paradigmas menos usuales como la selección por causalidad o los métodos aleatorios, se puede consultar [63] donde también se hace una revisión de las principales características de la selección de variables. Finalmente, otra referencia interesante puede ser [43], una revisión de los métodos de selección a través de los resultados del *NIPS 2003 Feature Selection Challenge*.

A estas alturas está clara la necesidad de reducir la dimensión y las ventajas generales que esto conlleva. Por ejemplo, debido a la maldición de la dimensionalidad es difícil clasificar un nuevo dato teniendo una muestra pequeña. Aun estando en un problema binario y teniendo m atributos binarios, el espacio de hipótesis mide 2^{2^m}. Cuanto más reducido el espacio de hipótesis, más sencillo encontrar la solución y además, el ratio de patrones por posibles soluciones aumenta exponencialmente con la reducción de dimensión mitigando el problema de la dimensionalidad [63]. Los objetivos concretos de las técnicas de selección de variables aparecen de forma muy similar en todas las referencias y según los lista [75] son:

- Eliminar variables inútiles o redundantes para mejorar el rendimiento temporal y el almacenamiento.
- Mejorar el acierto del clasificador disminuyendo el riesgo de sobreajuste.
- Hacer más comprensible el modelo.

Esto se consigue eligiendo un subconjunto adecuado de atributos desechando variables irrelevantes y/o redundantes. Dado el conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$, donde $x_i \in \mathbb{R}^D$ es un patrón con D atributos e y_i su etiqueta de clasificación, el objetivo de selección de variables consiste en tomar a lo sumo d características ($d \ll D$) para reducir el problema a un subespacio de $\mathbb{R}^d$. La dificultad radica en que el número de posibles subespacios de dimensión menor que d en $\mathbb{R}^D$ es $\sum_{i=1}^d \binom{D}{i}$, y para D grande (como es el caso) es imposible explorarlos todos en la práctica. Para solventar este inconveniente se han propuesto multitud de algoritmos basados en estrategias subóptimas. Estos métodos requieren un criterio para evaluar los distintos conjuntos y dirigir la búsqueda. Atendiendo a estos criterios se pueden hacer distintas clasificaciones de los selectores, pero la más estándar separa los métodos en tres ramas según su su manera de combinar la selección de variables con el modelo de clasificación: métodos de filtro, métodos *wrapper* y métodos embebidos. A continuación se esbozan todos ellos a partir de [42, 63] en general, junto con [89] para filtro, [58] en wrappers y el capítulo 4 de [43] para embebidos. Como referencias para los clasificadores que se van a citar se pueden utilizar las referencias de minería de datos [29, 12, 50].

Métodos de filtro Son aquellos en los que la selección de variables es independiente del clasificador ya que se realiza durante el preproceso. Típicamente se calcula una medida de relevancia y las variables con menor puntuación son eliminadas, mientras que el conjunto de las mejores

son presentadas al clasificador. Para ello sólo tienen en cuenta las propiedades intrínsecas de los datos. Esto los hace estadísticamente más robustos y eficientes, al requerir únicamente el cálculo del índice por el que se ordenarán las variables. Otras ventajas de los métodos de filtro es que tienen una buena escalabilidad, son computacionalmente sencillos y rápidos y la mencionada independencia del clasificador, lo que permite seleccionar una sola vez las variables y evaluar distintos métodos de clasificación. Esta misma independencia es una de las desventajas de estos métodos, en el sentido de que ignora cualquier posible interacción con el clasificador. Además los métodos de filtro son utilizados en ocasiones como primer paso de un selector en dos etapas, en el que realizan una primera selección de variables que es refinada después mediante un método de wrapper [52].

Métodos de wrapper Los métodos de wrapper son llamados así porque "envuelven" un clasificador. Este clasificador es utilizado a modo de caja negra para otorgar puntuaciones a las variables en función del acierto en clasificación. Para ello se realiza una doble búsqueda, primero en el espacio de variables donde se generan distintos subgrupos de atributos, que posteriormente son evaluados entrenando y testando el clasificador con ellos. Para converger al conjunto óptimo (o subóptimo) de variables es necesario el empleo de algoritmos de búsqueda, ya sean aleatorios o heurísticos. Todo este proceso es más complejo y costoso que en los algoritmos de filtrado (sobre todo si el entrenamiento del clasificador lo es), aumenta el riesgo de sobreajuste y depende enormemente del clasificador utilizado que debe ser el mismo que se vaya a utilizar en clasificación. Esto le hace perder poder de generalización pero puede explotar mejor las posibilidades del clasificador al aprovechar la interacción entre selector y clasificador y tienen en cuenta las dependencias entre variables de forma natural, cosa que muchos métodos de filtrado no. Algunas de las estrategias de búsqueda más comunes son las búsquedas secuenciales, de selección hacia delante (*forward*) o eliminación hacia atrás (*backward*) [57, 13, 104], el enfriamiento simulado (*simulated annealing*) [58] o los algoritmos genéticos [54]. Uno de los clasificadores más utilizados para estás técnicas, si no el que más, es SVM [60, 67].

Métodos embebidos Menos estudiados y utilizados de que los anteriores, en los métodos embebidos la selección de las variables se realiza en el proceso de entrenamiento del clasificador a partir del valor de una función objetivo a optimizar. Aprovechan mejor los datos y son más rápidos que los wrapper al no necesitar evaluar el clasificador para cada subconjunto de variables, pero son totalmente dependientes del clasificador. Se han utilizado en métodos como naive Bayes, árboles de decisión [27] y sobre todo SVM [103, 44].

Por su independencia del clasificador, facilidad para trabajar con grandes cantidades de atributos evitando el sobreajuste y mayor rapidez, se decidió utilizar un método de filtro. Sin embargo, hay

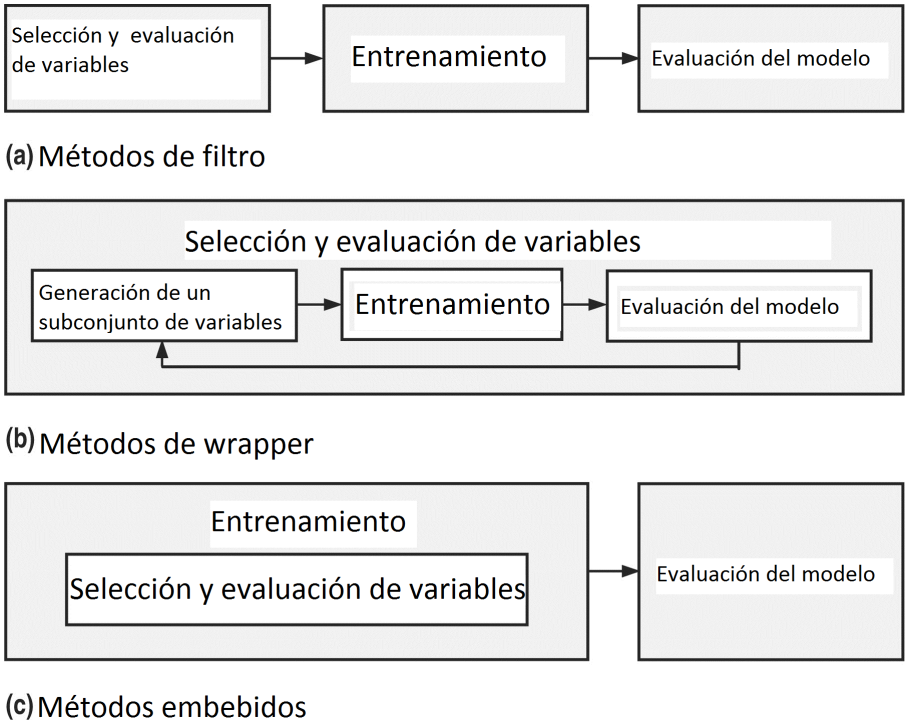

Figura 3.2.1: Esquema del funcionamiento de los tres paradigmas de selección de variables.

una gran variedad en función de la medida de relevancia, del enfoque univariado o multivariado o del tipo de búsqueda.

Para empezar, hay que eliminar las variables irrelevantes. Las redundantes no dejan de ser otro tipo de atributos irrelevantes [109]. Se diferencian en que una variable redundante implica la presencia de otra, ya que por si sola puede no ser irrelevante. Esta selección o eliminación de atributos puede llevarse a cabo de dos formas: ordenando las variables en función de un cierto criterio y tomando las k primeras (ranking), o seleccionando el conjunto que mejor resultado dé. El ranking suele corresponder con los métodos univariados. Estos métodos son más rápidos y menos costosos pero asumen una independencia entre las variables que no es real, de manera que no tienen en cuenta que variables poco relevantes por sí mismas pueden serlo en combinación de otras, ni que variables relevantes en solitario pueden ser redundantes con otras [43]. Ejemplos de estos algoritmos son la correlación de Pearson, la selección en función de test estadísticos sobre las variables como el T o el chi cuadrado [43], o la selección por ganancia de información [56].

Para cubrir esta carencia aparecen los métodos multivariados que tienen en cuenta las relaciones entre variables. Un ejemplo paradigmático de este tipo de algoritmos es el Relief y todos sus derivados [63, 43, 83], como ReliefF para problemas multiclase y datos incompletos y RRelief o

RReliefF para regresión. La idea de estos métodos aprovechando la filosofía de vecinos próximos es seleccionar aquellos atributos que resulten homogéneos entre los elementos más cercanos de una misma clase y distingan los representantes mas próximos de distintas clases. Otros ejemplos de este tipo de métodos son mRMR [28, 75] que selecciona la variable más relevante según la información mutua y va añadiendo iterativamente el atributo más relevante de los restantes y menos redundante con los ya seleccionados, y CFS (Correlation-based Feature Selection) que mide también la correlación entre pares de variables en lugar de sólo con la clase [45]. Se profundizará mRMR más adelante.

Además, todos los sistemas necesitan una estrategia de búsqueda, una medida de relevancia o calidad y un sistema de evaluación. Las metodologías de búsqueda son variadas y responden al hecho de que una búsqueda exhaustiva es computacionalmente imposible ya que para seleccionar d variables habría que explorar $\sum_{i=1}^{d} \binom{D}{i}$ conjuntos, lo que provocaría una complejidad $O(2^D)$. Por esto se han desarrollado múltiples estrategias que no garantizar alcanzar el óptimo, pero son computables y además no tienen el riesgo de sobreajuste de una búsqueda exhaustiva [63]. Destacan entre estas técnicas heurísticas la búsqueda hacia delante y hacia atrás, en las que cada paso se añade la variable más relevante o se elimina la más irrelevante respectivamente (en función del criterio elegido) [43]. Ejemplos representativos de estos procedimientos son la búsqueda secuencial hacia delante (SFS) y su análoga hacia atrás (SBS) [57]. Las búsquedas aleatorias representan una alternativa interesante [63] y otros métodos se encuentran en el capítulo 3 de [43].

Por su parte, la evaluación de un selector de atributos comprende dos aspectos principales: comparar los resultados antes y después de la reducción de variables para ver si se cumplen los objetivos de la selección, y comparar distintos métodos de selección entre sí. Se puede encontrar una guía de varios aspectos importantes a estudiar a la hora de evaluar un clasificador en [83], un ejemplo de comparativa entre distintos métodos en [82] y en [91] se estudian el tamaño y las relaciones entre los conjuntos de entrenamiento del clasificador y los utilizados para seleccionar las variables.

Finalmente, para que todo esto se pueda llevar a término hace falta definir cuáles son las variables relevantes o cómo medir la redundancia. En primer lugar se utilizaron medidas clásicas como la correlación de Pearson [78], estadísticos como el t o chi cuadrado [78], o la ganancia de información [56] utilizado también en árboles que compara la entropía del sistema antes de partir por un cierto atributo (o seleccionarlo) y después. Más tarde se añadieron medidas con otros enfoques o que mejoraban las anteriores como la información mutua [28], la medida basada en vecindad de Relief [83] o el estadístico de Fisher para selección de conjuntos [43]. Hay muchas más medidas, algunas específicas, por ejemplo para recuperación de información [37], otras empíricas como la

evaluación de curvas ROC [20]. Se hace un buen repaso de todas ellas en el capítulo 3 de [43]. Muchos de estos índices se utilizan en métodos de ranking de variables, pero como se ha dicho, esta aproximación puede quedar corta. En está línea se están investigando las interacciones entre variables [53, 112] para proponer indices que capturen las interacciones el de Relief, o se utilizan medidas de redundancia para combinarlas con las de relevancia en un nuevo indicador. Muchas de estas medidas son las mismas aplicadas a las variables en lugar de a las variables y a la clase, como sucede con la correlación (CFS) o la información mutua (mRMR).

Método de mínima Redundancia y Máxima Relevancia (mRMR)

Se ha visto una buena colección de métodos, técnicas y medidas de relevancia para la selección de variables, especialmente en cuanto a métodos de filtro se refiere, pero no hay un selector de variables superior en todos los casos, sino que cada algoritmo tiene sus propias características ajustándose mejor a ciertas aplicaciones concretas [63]. En este trabajo se ha decidido utilizar el mRMR [28, 75] ya que, a priori, sus características pueden lidiar bien con los datos funcionales (alta dimensionalidad y correlación), es bastante popular y accesible: en http://penglab.janelia.org/proj/mRMR/ puede utilizarse una versión online o descargarse otra más potente (también hay ya paquetes en programas como MATLAB con versiones del algoritmo). En cuanto a las características comentadas anteriormente mRMR es un método de filtro y por tanto, independiente del clasificador y con un buen rendimiento y escalabilidad. También es un algoritmo multivariado, que tiene en cuenta las relaciones entre variables, genera el conjunto óptimo mediante una búsqueda incremental y utiliza como criterio de relevancia la información mutua.

La información mutua es la medida más general de independencia estadística entre dos variables aleatorias, siendo capaz de cuantificar la información que estas comparten incluyendo las relaciones no lineales (punto al que no llega la correlación). Sean X e Y dos variables aleatorias continuas, $p(X)$ y $p(Y)$ sus funciones de densidad marginales, y $p(X,Y)$ su función de densidad conjunta, se define su información mutua como:

$$I(X,Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

Es fácil comprobar que así definida, la información mutua es siempre positiva (las funciones de densidad son ≥ 0 en todo punto) y que es simétrica, es decir, $I(X,Y) = I(Y,X)$. Y la información mutua será cero cuando las variables sean independientes y máxima cuando ambas tengan la misma distribución.

Como en la mayoría de los cálculos que incluyen funciones de densidad, el principal problema es estimar estas densidades a partir de los datos. Para variables discretas la ecuación se transforma

en el siguiente sumatorio:

$$I(X,Y) = \sum_i \sum_j P(x=i, y=j) \log \frac{P(x=i, y=j)}{P(x=i)P(y=j)}$$

pero en el caso continuo es mucho más costoso. Para solucionarlo se puede incorporar un paso de discretización de atributos en el preproceso, aunque esto añade el problema de determinar cuál es la discretización más adecuada [29].

Así, calculado la información mutua de cada conjunto de variables con la clase tendríamos el método de Máxima Dependencia [72]. Este es óptimo teóricamente, pero demasiado costoso en la práctica. Otra posibilidad es hacer un ranking en base a la información mutua de cada variable son la clase. Sin embargo, este criterio llamado de Máxima Relevancia [72] es univariado con los inconvenientes que conlleva. mRMR se convierte en multivariado al tener en cuenta, no sólo la relevancia de la variable que se selecciona, sino también la redundancia del conjunto de variables seleccionadas. Formalmente se define la relevancia de un conjunto de variables $S = X_1, ..., X_k$ con la clase C como:

$$V(S) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i, C)$$

y la redundancia del mismo conjunto de variables como:

$$W(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i, X_j)$$

Finalmente, para tener en cuenta las dos magnitudes maximizando la primera y minimizando la segunda, en [28] se propone utilizar un algoritmo iterativo incremental, en el que la primera variable seleccionada es aquella que tiene mayor relevancia y después se van añadiendo al conjunto de seleccionadas S la variable que maximice uno de los siguientes criterios:

- MID: El criterio de la diferencia es el más utilizado por su simplicidad y sus buenos resultados se elige la variable que haga máxima la diferencia $V - W$.
- MIQ: El criterio del cociente selecciona la variable que maximice V/W

En el artículo se define una formulación equivalente de estos criterios para variables continuas recurriendo al estadístico de Fisher. También se podría recurrir a técnicas de estimación de funciones de densidad no paramétricas bien conocidas, por ejemplo núcleos [29]. En general, y en los experimentos realizados, se opta por la discretización obteniendo buenos resultados [28].

Para finalizar, entre las nuevas aportaciones al método cabe citar la modificación introducida en [75] en la que se redefine la redundancia entre dos variables ponderando la información mutua por la entropía de la variable a añadir. De esta forma se consigue un nuevo operador no-simétrico que estima mejor la cantidad de información que añade la nueva variable en lugar de la cantidad de información que comparten. También merece una mención el nuevo método de selección de variables por programación cuadrática [84] propuesto como alternativa a mRMR. Este algoritmo ofrece resultados similares con un mejor rendimiento temporal en grandes conjuntos de datos gracias a la utilización del método de Nyström para diagonalización de matrices.

Capítulo 4

Clasificación funcional y selección de variables

Las motivaciones para tratar de aplicar los métodos de selección de variables a datos funcionales como paso previo a la clasificación supervisada, son principalmente tres: la falta de literatura al respecto, los buenos resultados obtenidos por estos métodos ante problemas similares y la interpretabilidad de los mismos.

La ausencia de bibliografía sobre este enfoque lo convierte en un caso de estudio interesante y novedoso. Es llamativa la falta de estudio de este tipo de métodos en el ámbito funcional, ya que a priori lidian con dos de los problemas fundamentales de las funciones, la alta dimensionalidad y la correlación entre puntos. En el caso de la clasificación no-supervisada se pueden encontrar algunas aproximaciones como la de Fraiman et al. [39], si bien escasas y tratando únicamente de eliminar variables ruidosas o redundantes, mientras que en el nuevo enfoque propuesto se tiene en cuenta también la relevancia de la variable para la clasificación. Por su parte, en la clasificación supervisada, los únicos trabajos relacionados tratan de conseguir métodos de proyección más interpretables [92], o realizan preprocesos de las funciones para extraer una cierta información que no pueden ser denominados selección de variables. Este sería el caso de [61] donde se tomarían unas nuevas variables generadas a partir de ciertos segmentos de la curva.

No hay mucho más que decir sobre la eficacia de los algoritmos de selección de variables, se pueden encontrar referencias en la sección anterior (3.2, página 41). Pero la interpretabilidad de estos métodos es un factor a tener en cuenta. Hasta ahora, los métodos aplicados a curvas para reducir la dimensión son poco interpretables [35] pese a algunos recientes intentos [92]. Sin embargo, las funciones y en concreto las curvas, tienen una profunda componente visual y son fáciles de interpretar. Para cualquier persona es relativamente sencillo diferenciar un año frío de uno cálido, o un

periodo positivo para la bolsa de uno negativo a través de sus correspondientes gráficas. Es decir, clasificamos de forma natural y con cierta facilidad las funciones sólo por su "aspecto". Además, para hacerlo no tenemos en cuenta toda la curva, sino que nos fijamos en puntos concretos. Por ejemplo, un médico, a la hora de clasificar un electrocardiograma tiene en cuenta entre 5 y 9 variables de toda la curva del electro (ver 4.0.2). Del mismo modo, un selector de variables se quedaría con ciertos puntos de la función. Éstos tendrían su significado, facilitarían la interpretación del modelo por parte de expertos para desarrollar nuevas conclusiones y en caso de estar bien diseñado, además mejoraría la clasificación automática. Veremos a continuación que mRMR consigue buenos resultados y selecciona variables con sentido "visual".

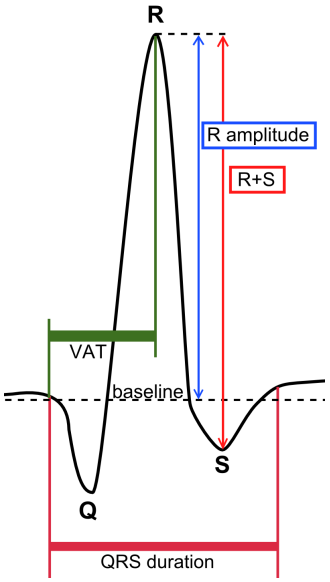

Figura 4.0.2: Algunas de las variables que se consideran para estudiar un electrocardiograma. No se mira toda la curva sino puntos concretos.

4.1. Experimentos

El objetivo principal de esta sección es aplicar mRMR en varios conjuntos de datos funcionales y ver si es un punto de partida válido para futuros estudios. Por tanto, al ser la base de un trabajo en curso, no pretende ser exhaustiva sino dar una visión general del comportamiento y capacidades del método y su viabilidad. Para ello, además de medir errores de clasificación, se compara con métodos de selección univariados a través de MaxRel, con algoritmos de proyección de probada eficacia mediante PLS, e incluso se hacen algunas observaciones sobre los clasificadores.

En primer lugar se comentarán algunos detalles de implementación de los distintos algoritmos de reducción de dimensiones y de los clasificadores utilizados, para continuar con la descripción de

los conjuntos de datos utilizados. Finalmente, se expondrán y analizarán los resultados obtenidos.

4.1.1. Implementación

Para la realización de las pruebas fue necesario implementar el método de Mínima Redundancia y Máxima Relevancia junto con el de Máxima Relevancia y los Mínimos Cuadrados Parciales para compararlos. Además, hubo que implementar el clasificador. Aunque mRMR es un selector de filtro, y teóricamente es independiente del clasificador, se decidió realizar pruebas con dos clasificadores, SVM y KNN.

Por simplicidad, se decidió utilizar el código de mRMR disponible en la web de uno de los autores, http://penglab.janelia.org/proj/mRMR/. Este es un código testado, disponible en distintos lenguajes y para distintos entornos. El problema de las variables continuas se resuelve con una discretización de los atributos, solución sencilla y de bajo coste que no empeora los resultados [28]. La web también ofrece una versión simplificada on-line con algunas opciones básicas para conjuntos reducidos. Además, este programa genera el ranking de Máxima Relevancia.

Por las mismas razones, para el SVM se ha utilizado el conocido paquete LIBSVM en su versión 2.8 [18]. Este paquete es válido para problemas binarios y multiclase, y permite la utilización de distintos núcleos. En este caso se utilizaron el lineal y el gaussiano.

El resto del código se implementó en MATLAB. Esto podría ser un inconveniente para llevar a cabo pruebas exhaustivas, pero no entra en conflicto con los objetivos de este trabajo. En este lenguaje se programó un algoritmo de K vecinos próximos estándar [5] con la distancia euclídea y dos versiones de PLS. La primera versión está basada en el algoritmo iterativo NIPALS propuesto por Wold. Se eligió está versión iterativa para problemas binarios debido a los problemas del enfoque de valores propios (3.1, página 38). La versión implementada corresponde al citado como modelo PLS1 en [98]. Por otra parte, para problemas multiclase, se implementó la solución del problema de valores propios [65].

4.1.2. Conjuntos de Datos

Se han elegido 4 conjuntos de datos funcionales para este trabajo, uno de ellos se ha trabajado en crudo y diferenciado. Todos ellos son datos reales utilizados en publicaciones. Tres son binarios, recogiendo la problemática básica, y uno multiclase. Algunas de sus características se recogen en la tabla 4.1.2.

Muestra	N	D	C	Acierto de referencia	Referencias
ECG	2026	85	2	99.26 %	[4, 5, 101]
Yoga	306	426	2	95.24 %	[101, 102]
Tecator	215	100	2	76.76 %	[4, 5, 35, 61]
Tecatordif	215	99	2	96.32 %	[35, 61]
Phoneme	4509*	256	5	89.80 %	[4, 5, 35, 49, 61]

Cuadro 4.1.1: Descripción de los conjuntos de datos. N representa el número de muestras y D su dimensión. C es el número de clases del problema. El acierto de referencia es el calculado con todas las variables con 3NN en los casos binarios y 9NN en los fonemas. Se incluyen referencias a trabajos donde se tratan estos conjuntos en clasificación. * El conjunto contiene 4509, en los experimentos se ha reducido la muestra a 1000

ECG Conjunto de datos del MIT-BIH Arrhythmia Database (*http://www.physionet.org*). Están disponibles en la página *http://www.cs.ucr.edu/ẽamonn/time_series_data/*, que contiene un banco de datos de series temporales. Cada muestra recoge la señal de un electrodo durante un latido y consta de 2026 muestras, 1506 normales y 520 anormales. Todas las curvas están normalizadas y reescaladas para tener longitud 85.

Yoga Formado por 316 muestras de longitud 426. La muestra captura la transición entre dos posturas de yoga. Se diferencian dos clases en función de si la persona que hace el cambio de postura es un hombre (150) o una mujer (156). Se puede encontrar en la base de datos de series temporales *http://www.cs.ucr.edu/ẽamonn/time_series_data/*.

Tecator Compuesto por 215 curvas de espectros de absorbencia de 100 canales(dimensión) de sendas muestras de carne. Pueden encontrarse en *http://lib.stat.cmu.edu/datasets/tecator*. Estos datos tienen también valores de humedad, grasa y proteínas. Para clasificarlos se dividen en dos grupos según tengan menos de un 20 % de grasa o no [35]. Una particularidad de este conjunto es que los datos en crudo son muy homogéneos y difíciles de clasificar [35, 61], por lo que se decidió seguir el ejemplo de las referencias y estudiar también el conjunto diferenciado (ver 4.1.1).

Phoneme Conjunto de datos de reconocimiento locutor. Inicialmente introducidos en [49], los datos y su descripción están en la página web de [50], *http:www-stat.stanford.edu/ElemStatLearn*. El conjunto lo forman 4509 muestras que recogen la grabación de uno de los siguientes 5 fonemas

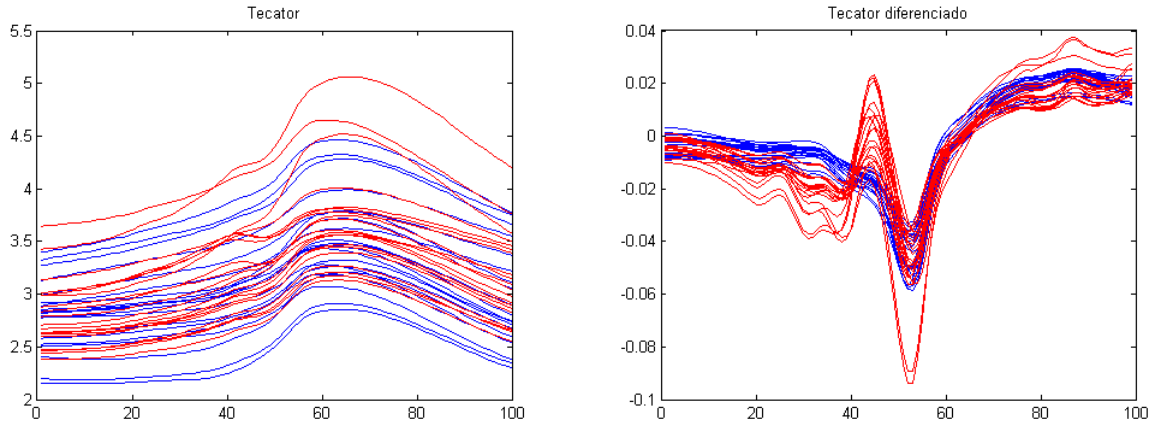

Figura 4.1.1: Muestra de 15 curvas de cada clase del conjunto Tecator, a la izquierda en crudo y a la derecha tras diferenciarlas. El conjunto diferenciado es claramente más fácil de clasificar.

(clases): “sh”(872), “dcl”(757), “iy”(1163), “ao”(1022) y “aa”(695). Los ejemplos están discretizados a dimensión 256. Como este conjunto resultaba poco manejable en MATLAB, se seleccionaron aleatoriamente 200 muestras de cada clase para los experimentos.

4.1.3. Metodología y resultados

No hay una metodología estándar en las referencias estudiadas. Se realizan muestreos de distintos tamaños con distinto número de permutaciones [4, 34, 61], incluso en [101] se utiliza el mismo conjunto para entrenamiento y test. Ante esta heterogeneidad, se optó por un método de validación cruzada de 10 hojas [29] con 100 ejecuciones para estimar la tasa de acierto de los métodos. De este modo se consiguen unas estimaciones consistentes y que pueden compararse con la mayoría de la bibliografía desde una cierta cautela.

Las pruebas se encaminaron a esbozar los puntos comentados anteriormente. Así, se realizaron experimentos con los dos clasificadores. Una vez fijado el clasificador se probó la eficacia de mRMR de selección en función del acierto base y en comparación con PLS y MaxRel.

4.1.3.1. Sobre los clasificadores

Por cuestiones de claridad se decidió hacer el análisis de mRMR sobre un único clasificador, ya que además, al tratarse de un método de filtro debería ser relativamente independiente. Sin embargo, se prefirió hacer unas pruebas iniciales con dos modelos representativos y con una fuerte presencia en la literatura, SVM [61, 87, 99] y KNN [3, 35]. En SVM se probó con un núcleo lineal y uno gaussiano, ambos con las opciones por defecto de LIBSVM [18], mientras que para KNN se eligió $K = 3$ para problemas binarios y $K = 9$ para el de los fonemas debido a su mayor complejidad.

Los resultados presentados en la tabla 4.1.3.1 muestran que no hay una clara superioridad, aunque KNN tiene en general mejor comportamiento general hay excepciones, y hay una cierta igualdad en cuanto a los máximos. Además, KNN parece claramente superior en las primeras etapas cuando la reducción es mayor. Por su parte, SVM se muestra más débil con el núcleo lineal, tardando más en arrancar, aunque en los problemas binarios acaba superando al gaussiano que se desenvuelve mejor en los espacios más reducidos. Un ejemplo de esta superioridad es en la primera variable de ECG debido a que una clase queda rodeada por la otra en ese punto (ver figura 4.1.2). Para obtener mayores conclusiones habría que hacer un estudio más exhaustivo de los parámetros (SVM es más complejo y tiene más parámetros que ajustar), pero parece que se cumplen las premisas de los métodos de filtro [43] y se tienen comportamientos similares.

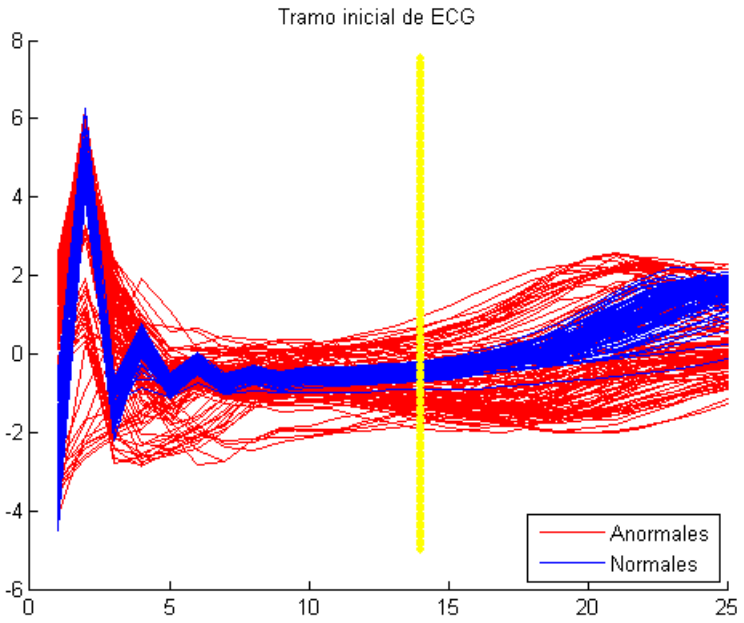

Figura 4.1.2: La línea amarilla representa la primera variable elegida por mRMR, en este caso la 14. En este punto las curvas de los electrocardiogramas anómalos rodean casi perfectamente a las de los normales. Esto es lo que hace que el núcleo gaussiano funcione tan bien.

Por otra parte, KNN es propuesto en varios artículos como posible método de prueba general por sus buenas cualidades [3]. En esta línea, Hand [47] aboga por métodos simples (menos costosos y más interpretables) arguyendo una falsa ilusión de progreso en los algoritmos más sofisticados. Por todas estas razones, por el menor coste computacional de KNN, y buen comportamiento en los ejemplos frente a SVM, se eligió el método de vecinos próximos con $K = 3$ para los problemas binarios y $K = 9$ para el multiclase.

	ECG			**Yoga**			**Phoneme**		
#Variables	3NN	SVMl	SVMg	3NN	SVMl	SVMg	9NN	SVMl	SVMg
1	92.70 %	88.73 %	94.65 %	86.77 %	85.12 %	89.08 %	56.29 %	56.56 %	55.85 %
2	95.84 %	82.81 %	83.96 %	92.63 %	86.04 %	92.91 %	78.18 %	71.09 %	76.77 %
3	96.10 %	79.61 %	81.09 %	93.24 %	88.03 %	95.02 %	84.93 %	74.02 %	78.63 %
4	95.74 %	80.51 %	81.41 %	95.06 %	91.87 %	94.96 %	85.03 %	74.37 %	79.92 %
5	96.55 %	86.02 %	88.32 %	94.69 %	90.00 %	95.12 %	86.96 %	81.95 %	86.87 %
10	99.59 %	96.55 %	97.04 %	95.49 %	94.15 %	93.54 %	90.49 %	89.47 %	90.07 %
20	99.64 %	97.86 %	97.12 %	95.38 %	95.07 %	93.49 %	91.27 %	90.60 %	90.78 %
Total	99.25 %	98.19 %	94.98 %	95.24 %	94.23 %	91.17 %	89.80 %	91.26 %	92.73 %

Cuadro 4.1.2: Resultados de KNN frente a SVM. En la columna izquierda se muestra el número de variables seleccionadas con mRMR para los que se obtuvieron los resultados. En Total se sitúa el acierto del clasificador con todas las variables del problema. SVMl hace referencia a SVM con el núcleo lineal y SVMg al gaussiano.

4.1.3.2. mRMR, MaxRel y PLS

Además de fijar el clasificador, de la tabla 4.1.3.1 también se desprende que la utilización de mRMR es viable en este tipo de conjuntos ya que con 10 variables o menos, se supera el nivel de acierto tomado como referencia con todas las variables. Una vez comprobado este punto y fijado el clasificador, los siguientes experimentos están encaminados a comprobar si la mejora teórica de mRMR sobre MaxRel es real, y si el método de selección de variables es competitivo con los métodos de extracción de características ya asentados en FDA, o no aporta nada. Para esta segunda comparación se medirá con PLS como representante de los algoritmos de reducción por proyección, que además está especialmente indicado para clasificación supervisada [65]. Asimismo, se pondrán de manifiesto algunas de las características visuales y de interpretabilidad de mRMR comentadas en secciones anteriores.

Los resultados obtenidos en los experimentos se exponen en la tabla 4.1.3.2. La primera conclusión que se puede extraer de los mismos es que mRMR mejora sustancialmente a MaxRel, si bien puede parecer menos clara en los casos binarios al partir de una primera variable muy relevante (la primera coincide en los dos métodos). La única excepción es el conjunto Tecator diferenciado, ya que posee dos variables muy relevantes y similares que pueden explicar todo el conjunto y son elegidas al inicio en ambos métodos. MaxRel [72] puede verse como un precursor de mRMR ya

	ECG			**Yoga**			**Tecator**			**Tecatordif**			**Phoneme**		
#Variables	MaxRel	mRMR	PLS	MaxRel	mRMR	PLS	MaxRel	mRMR	PLS	MaxRel	mRMR	PLS	MaxRel	mRMR	PLS_2
1	92.71 %	92.7 %	82.16 %	85.95 %	86.77 %	82.24 %	65.41 %	65.74 %	61.82 %	95.27 %	95.8 %	83.32 %	55.99 %	56.29 %	59.27 %
2	93.71 %	95.85 %	94.31 %	87.9 %	92.63 %	91.16 %	65 %	64.04 %	74.92 %	97.23 %	97.24 %	91.53 %	62.58 %	78.18 %	76.06 %
3	95.33 %	96.11 %	96.36 %	88 %	93.24 %	92.25 %	64.19 %	67.49 %	76.62 %	97.09 %	96.45 %	96.12 %	66.08 %	84.93 %	86.19 %
4	95.72 %	95.75 %	97.7 %	87.81 %	95.06 %	93.13 %	64.93 %	78.35 %	76.74 %	96.96 %	96.23 %	96.77 %	68.02 %	85.03 %	90.14 %
5	95.97 %	96.56 %	99.14 %	87.76 %	94.69 %	94.84 %	64.46 %	78.09 %	76.7 %	96.62 %	96.88 %	96.57 %	71.42 %	86.96 %	90.21 %
10	97.24 %	99.59 %	99.6 %	92.57 %	95.49 %	96.27 %	67.77 %	75.11 %	76.85 %	96.22 %	96.23 %	96.51 %	74.55 %	90.49 %	90.24 %
20	99.48 %	99.65 %	99.65 %	92.71 %	95.38 %	96.29 %	70.68 %	77.59 %	76.85 %	95.54 %	95.61 %	96.08 %	85.11 %	91.27 %	90.59 %
Máx	99.77 %	99.77 %	99.71 %	95.31 %	96.05 %	96.33 %	76.42 %	78.97 %	76.85 %	97.23 %	97.24 %	96.77 %	91.9 %	91.93 %	91.08 %
	(34)	(33)	(75)	(123)	(36)	(12)	(89)	(35)	(7)	(2)	(2)	(4)	(145)	(99)	(62)
Base	99.26 %			95.24 %			76.76 %			96.32 %			89.8 %		

Cuadro 4.1.3: Resultados de mRMR frente a MaxRel y PLS. En la primera columna se muestra el número de variables seleccionadas, y en las restantes los resultados obtenidos en clasificación con un 9NN en el conjunto Phoneme y 3NN en el resto. En Base se sitúa el acierto base para el problema, y en Máx el mayor porcentaje de acierto obtenido y el número de dimensiones en que se obtiene entre paréntesis. PLS_2 hace referencia a la formulación matricial de PLS, mientras que con PLS se indica el modelo iterativo

que utiliza la información mutua con la clase para realizar un ranking de atributos por relevancia. El simple hecho de utilizar información mutua como medida de dependencia estadística en lugar de una medida lineal como la correlación anula la ventaja de mRMR por el uso de esta métrica, pero el inconveniente para MaxRel, es que los datos funcionales tienen una correlación muy alta, con lo que toma una franja de variables muy redundantes antes de pasar a otro punto de interés (ver figura 4.1.3). Estas franjas, aun siendo de variables muy relevantes no aportan apenas información nueva y ralentizan y entorpecen la clasificación. Por su parte, en la misma figura se puede apreciar la distinta filosofía de mRMR que tras seleccionar una variable relevante busca la siguiente de forma que no sea redundante con las anteriores. Así, en lugar de franjas selecciona variables sueltas en distintos puntos de la función. Este comportamiento ilustrado con el conjunto ECG se repite en todos los demás.

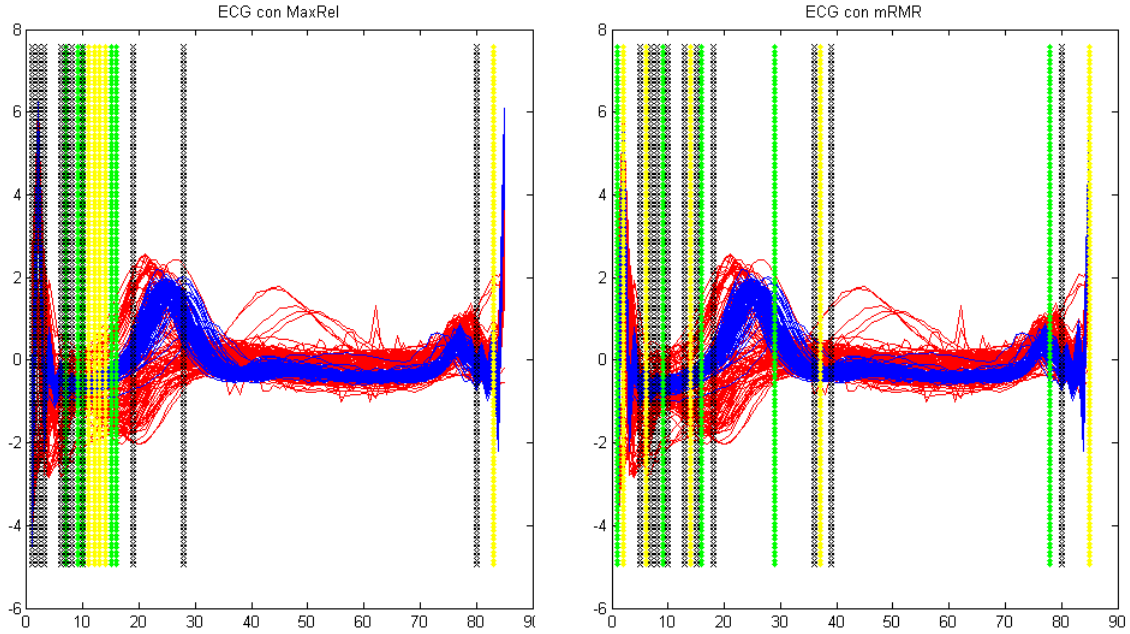

Figura 4.1.3: Ambas gráficas muestran un subconjunto de las curvas del conjunto ECG. Las rayas verticales indican las primeras 20 variables seleccionadas por MaxRel (a la izquierda) y mRMR (a la derecha). Para tener una idea de la secuencia con la que son añadidas, las 5 primeras variables se pintan en amarillo, las 5 siguientes en verde, y las restantes en negro. En las gráficas se observa la selección de un rango casi completo en MaxRel y la mayor variedad recogida por mRMR

La segunda conclusión importante es que un método de selección de variables como mRMR es, a priori, plenamente competitivo con los métodos de extracción de características, uno de cuyos representantes más destacados es PLS. En los resultados expuestos en 4.1.3.2, mRMR se muestra

al nivel de PLS mejorándolo en muchos casos. Especialmente, el selector mejora a su rival en los problemas binarios y en las dimensiones más bajas, y PLS converge a máximos antes que mRMR (especialmente en el problema multiclase), si bien están bastante parejos. Podría pensarse que la diferencia en el conjunto Phoneme fuera producida por la distinta implementación, pero no es así, ya que si se hubiera implementado la versión de valores propios de PLS [6] en el caso binario sólo habría generado una componente con autovalor distinto de cero, y se puede comprobar que ésta es la primera que genera el método iterativo. También se aprecia una ralentización, y en ocasiones empeoramiento, de mRMR después de las primeras componentes. Esto puede deberse a que PLS va reconstruyendo la "información" total del espacio constantemente y no empeora (cada nueva componente es ortogonal). Sin embargo, la selección de variables añade variables nuevas que pueden repetir información o incluso mezclar los datos (la nueva variable se ha elegido por una búsqueda subóptima). Por este mismo motivo, la estabilización de PLS es mucho más rápida que la de mRMR, que aunque alcanza mejores resultados tiene una mayor variabilidad. Esto se refleja con más intensidad en el caso de Tecator, en el que la uniformidad de las curvas "confunde" al selector. Estas consideraciones se pueden apreciar en la gráfica 4.1.4. En cualquier caso, ambos métodos superan la tasa de acierto base y con pocas componentes, mostrándose que puede tener mucho sentido la aplicación y estudio de los métodos de selección de variables en FDA. Además, es llamativo que en todos los casos excepto Tecator, con dos variables mRMR mejore a PLS, y que con uno tenga mejores resultados en los problemas binarios. Esto parece confirmar la intuición de que las curvas pueden discriminarse de forma eficiente mirando sólo algunos puntos concretos, sin tener que recurrir a variables sintéticas o a valorar el conjunto. Así lo pone de manifiesto el hecho de que dos variables propias de la función (que además tienen un significado) separen mejor el conjunto que las dos variables sintéticas generadas por PLS. Algunos ejemplos pueden verse en la figura 4.1.5.

En cuanto a la representatividad y la potencia visual del método ya se han mencionado dos puntos a favor. En primer lugar, se ha visto que la elección de variables es "inteligente" en el sentido de que evita la redundancia y no selecciona rangos (figura 4.1.3). En segundo lugar, a través de la figura 4.1.5, se observa que la selección es representativa y genera un subespacio con una buena separación de las clases. Si pintamos las primeras componentes enfrentadas dos a dos en el problema de Phoneme se ve cómo el algoritmo parece seguir el criterio natural de ir separando las clases una a una. De este modo, en la figura 4.1.6 se ve como las dos primeras variables separan tres de los fonemas, y la primera y la tercera distinguen entre los otros dos. En tercer lugar, la selección parece concordar en muchos casos con la elección "a ojo" que haríamos para clasificar las curvas (ver figura 4.1.7). Esto sugiere que este método, u otros que se pudieran desarrollar, podrían emular con ciertas garantías el conocimiento experto. Pero además dan un paso

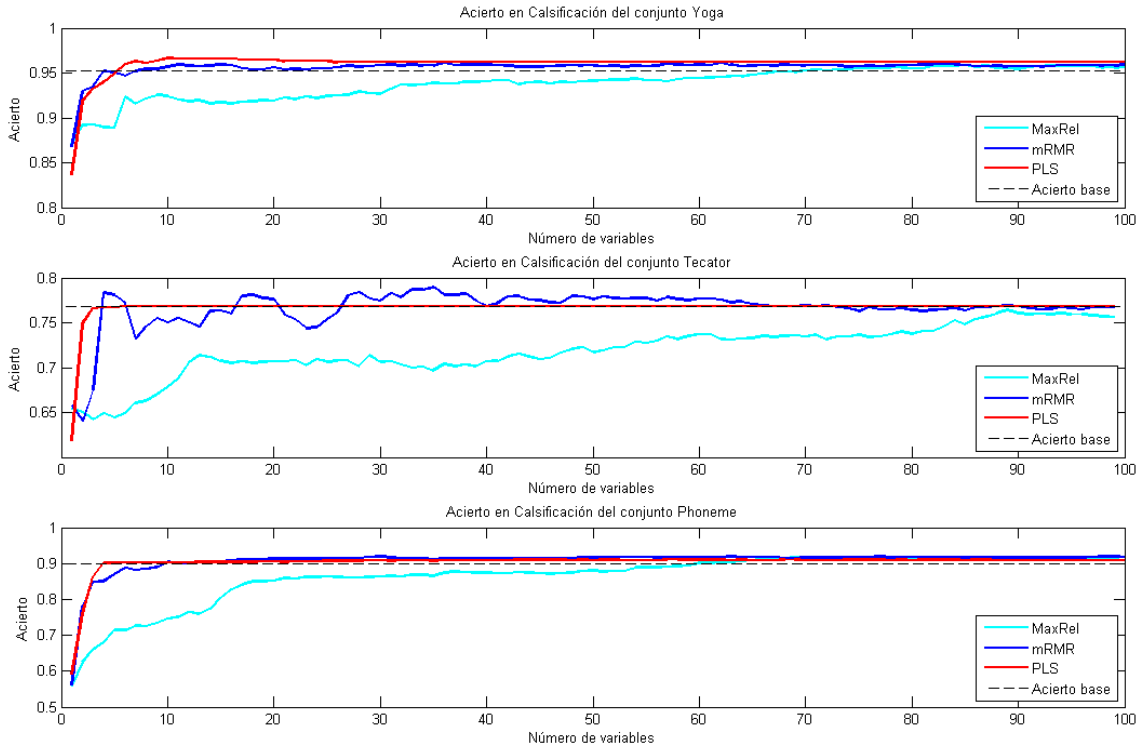

Figura 4.1.4: Las gráficas muestran la evolución para las primeras componentes de los tres métodos de reducción de dimensión analizados para los conjuntos de Yoga (arriba), Tecator (centro) y Phoneme (abajo). En las tres gráficas se puede ver la superioridad de PLS(rojo) y mRMR(azul oscuro) sobre MaxRel(azul claro), y la igualdad o leve superioridad de mRMR sobre PLS en el arranque y en los máximos. Asimismo, se aprecia la pronta estabilización de PLS y el resto de características comentadas. El acierto base está representado por la línea discontinua.

más allá e incluso cuando seleccionan variables que nos pasarían más desapercibidas, muchas veces se puede ver un significado claro. Por ejemplo, en el caso de ECG nos podría parecer mejor elegir una variable en torno a la 26 en lugar de las que elige el método. Sin embargo, si observamos las medias y las varianzas de los datos, vemos que las primeras variables elegidas corresponden a una gran diferencia en las medias de los datos o a variables con una ligera separación pero con al menos una de las clases muy concentrada (con muy poca varianza) como se observa en la figura 4.1.8. El intentar que las clases estén separadas entre sí y lo más agrupadas posible es un criterio clásico en clasificación, utilizado por ejemplo en Fisher. Prueba de esta buena elección es que si tomamos la variable 26, obtenemos un acierto de 76.73 % frente al 92.7 % obtenido con la 14 que selecciona mRMR. A todo esto hay que añadir que las variables seleccionadas son, en general, muy interpretables. Pueden representar el canal de un espectro, una fecha y hora en un índice económico, un determinado momento de la emisión, o el instante concreto del latido, lo que facilitaría el trabajo

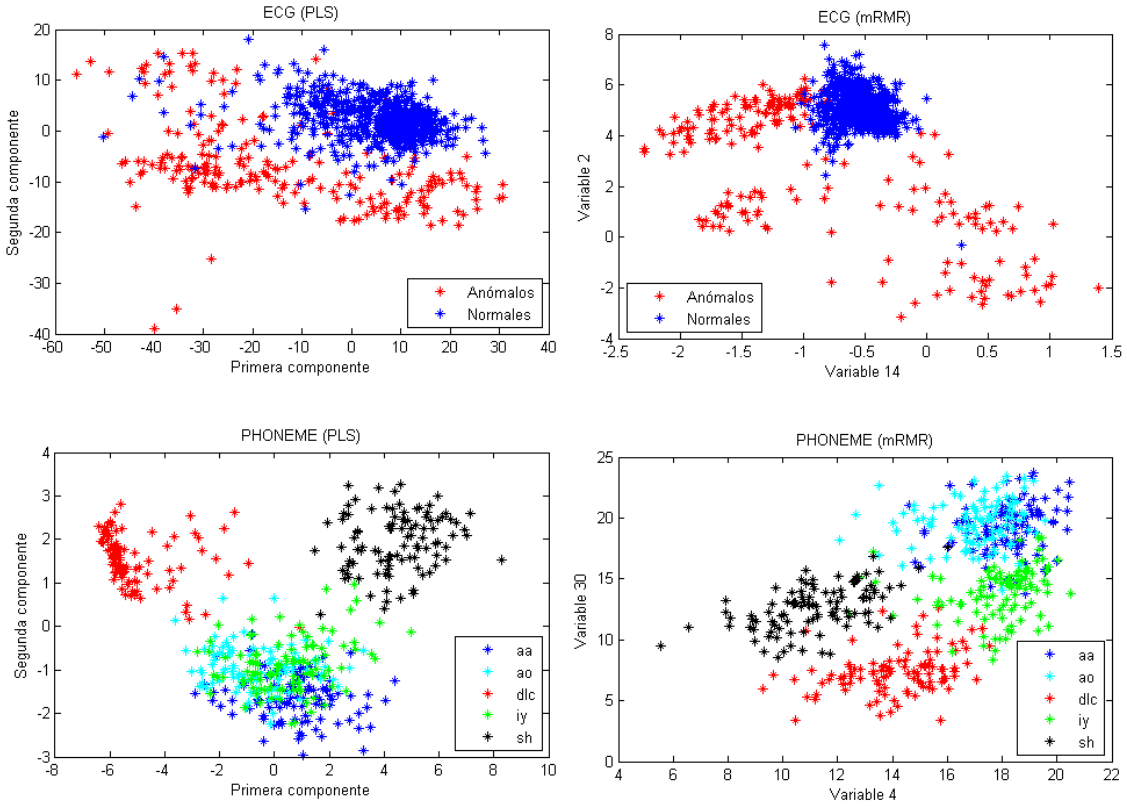

Figura 4.1.5: Representación de los datos de ECG en la parte superior y del conjunto Phoneme reducidos a dimensión 2 con ambos métodos. En la parte izquierda se pueden ver el resultado con las componentes generadas por PLS y en la derecha el análogo con las variables seleccionadas con mRMR. Visualmente parece mucho más apropiado el resultado de mRMR ya que en el primer conjunto lo separa más, y en el segundo separa tres fonemas mientras que PLS sólo logra separar dos.

posterior de los expertos a la hora de sacar conclusiones y aporta un valor extra del que carecen los métodos de extracción de características.

Además, a partir de los datos y durante las pruebas se han podido observar una serie de conclusiones menores para los objetivos de este trabajo, pero que refuerzan las ideas esbozadas sobre análisis funcional. Aparece la importancia del preprocesado de las curvas en problemas complejos y la utilidad de la diferenciación [81] se pone de manifiesto en el caso de Tecator. También queda patente la relevancia del ajuste de los parámetros del modelo de clasificación [35]. Aunque no se ha hecho especial hincapié, sí se probaron distintas configuraciones de parámetros y, por ejemplo, para tratar con problemas más complejos como Phoneme fue necesario incrementar la K de KNN. Y también se ponen de manifiesto, entre otros, los dos grandes problemas de los datos funcionales,

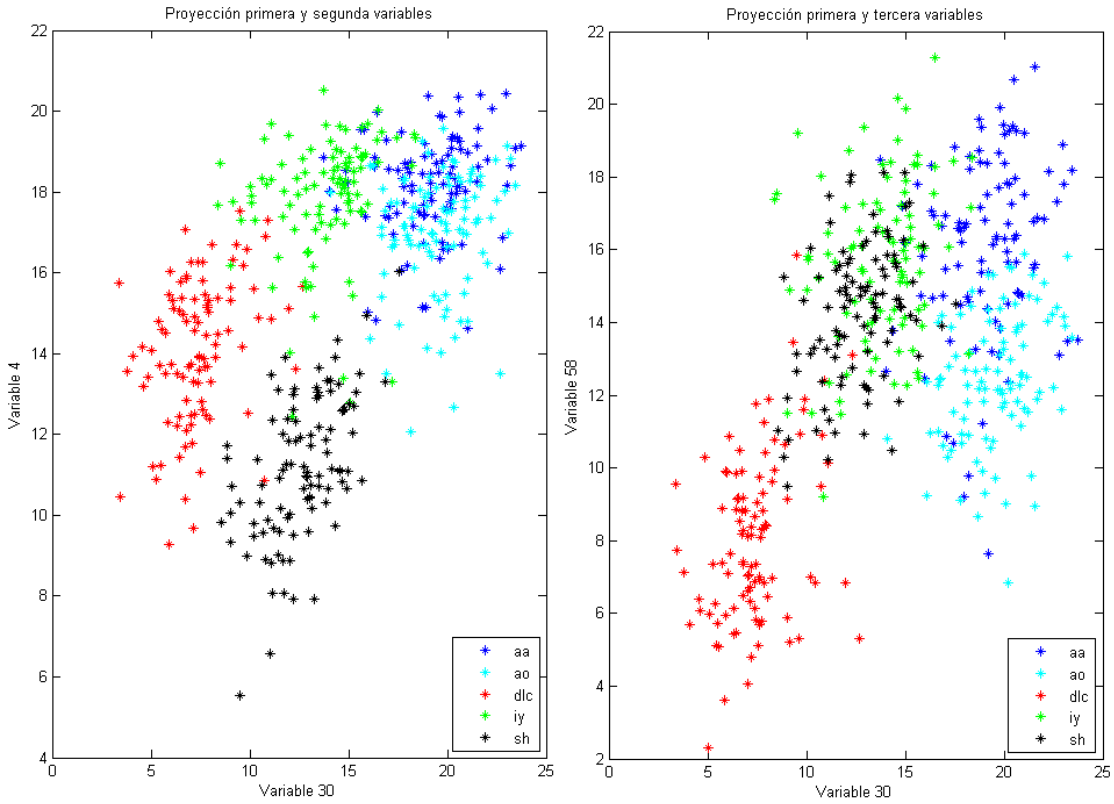

Figura 4.1.6: Disposición de los datos del conjunto Phoneme considerado las tres primeras variables de mRMR. A la izquierda las dos primeras y a la derecha la primera con la tercera. Aunque no lo consiguen plenamente, la tercera variable complementa a las dos primeras tratando de separar las clases que quedan mezcladas con las dos primeras.

la alta dimensionalidad y la elevada correlación, que han condicionado el trabajo (para reducir la primera) y los resultados.

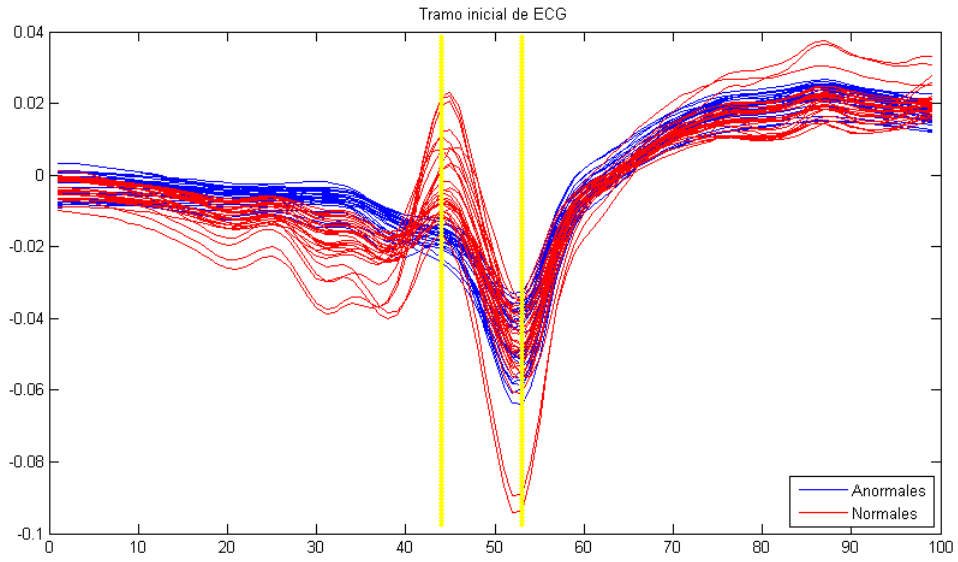

Figura 4.1.7: El la gráfica se muestran las curvas de Tecator diferenciado y las dos primeras variables que selecciona mRMR en amarillo. Con estas dos variables se consiguen los mejores resultados en clasificación y coinciden con los puntos que un observador escogería a la vista de la gráfica para clasificar. Otra pareja de variables equivalente a estos efectos se obtendría sustituyendo la 53 por la 36. Aunque no se ha pintado por claridad, dependiendo del conjunto de entrenamiento, ambas variables intercambian sus posiciones, lo que indica que mRMR se centra en los tres puntos de interés que percibimos y que dan los mejores resultados.

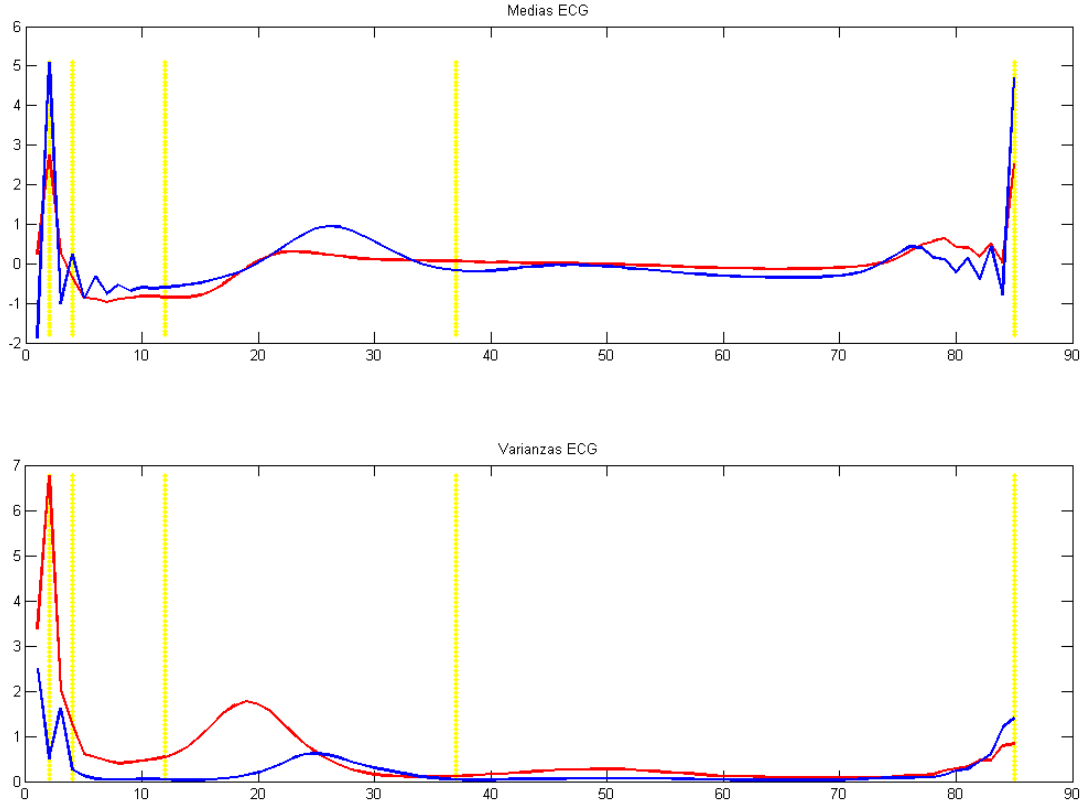

Figura 4.1.8: Se representan las 5 primeras variables de mRMR sobre los gráficos de las medias de las clases de ECG (arriba) y sus varianzas (abajo). Se aprecia que las variables escogidas separan bien las medias de las clases (primera y última) o las separan menos pero eligen puntos donde las varianzas son muy pequeñas.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones generales

El análisis de datos funcionales es una rama de la estadística muy extensa y compleja que aborda gran variedad de problemas y aglutina distintos enfoques dentro de ella. Un estudio detallado es inasumible en un trabajo de esta índole, pero se ha realizado un esfuerzo de síntesis para dar una visión general y actual de las características, metodologías y problemática del área, poniendo el foco en la clasificación supervisada. De la revisión bibliográfica se deduce que FDA es un campo de gran interés y en expansión, como muestra el creciente número de publicaciones. La importancia del análisis de datos funcionales radica en gran medida en el valor y alcance de los indicadores o magnitudes con naturaleza funcional. Estas señales se encuentran cada vez más en áreas de gran importancia como la medicina [55], la economía [88] o el reconocimiento biométrico (seguridad) [105]. Además, FDA es un campo de investigación relativamente reciente debido a que se ha necesitado un gran avance tecnológico permite captar con garantías y poder manejar este tipo de datos. Por ello todavía existe mucho trabajo pendiente, tanto para seguir profundizando en aspectos teóricos y computacionales, como para explorar nuevas vías [35, 100]. En especial, se pueden aplicar todas estas consideraciones a la clasificación supervisada por la importancia de discriminar entre curvas o imágenes.

Otra conclusión observada durante la realización del trabajo es la "falta de comunicación" entre distintas disciplinas. Al estudiar las referencias de FDA y las que podríamos denominar de minería de datos, se ha visto como dos áreas con muchos puntos y objetivos comunes difieren enormemente en algunos aspectos hasta el punto de que metodologías muy asentadas en una de ellas, no son conocidas o no tienen su contrapunto en la otra, como es el caso de la selección de variables. Una parte de las diferencias se justifica en los distintos enfoques, sin embargo, parecería más productivo una mayor comunicación.

Los experimentos realizados han puesto de manifiesto los efectos de la maldición de la dimensionalidad en los datos funcionales y la necesidad de reducir su dimensión como paso previo a la clasificación. Tanto mRMR como PLS mejoran el acierto base obtenido con todas las variables. En algunos casos la mejora es sustancial, con aumentos de la tasa de acierto superiores al 2 % en cifras ya bastante elevadas. Además, son necesarias muy pocas variables para llegar a niveles de acierto cercanos al máximo, generalmente menos de 10 (en Tecatordif se alcanza el máximo con sólo 2 variables en mRMR y 4 en PLS), consiguiendo una reducción de unos dos órdenes de magnitud en el número de atributos con las consiguientes ventajas computacionales y de almacenamiento.

Finalmente, el estudio sobre la viabilidad de la utilización de técnicas de selección de variables en datos funcionales, uno de los objetivos del trabajo, desprende varias conclusiones. Por un lado, los métodos univariados no parecen una alternativa razonable debido a que asumen independencia entre las variables cuando las funciones presentan una altísima correlación. Prueba de ello son los pobres resultados obtenidos por el método de Máxima Relevancia que es claramente superado por PLS. Sin embargo, el comportamiento de mRMR es bastante bueno. Su combinación de premiar la relevancia con la clase y penalizar la redundancia entre variables parece ajustarse bien a los datos funcionales y consigue buenos resultados de forma eficiente, mejorando en la mayoría de los casos a PLS. Además de obtener buenos niveles de acierto, se ha visto que un método como mRMR aporta un grado de interpretabilidad del que carecen las variables sintéticas de los algoritmos de extracción de características como PLS. Por todo ello, si bien se trata de un estudio preliminar, se puede concluir que la filosofía de los métodos de filtro multivariados como mRMR es perfectamente aplicable a datos funcionales y abre la puerta a un mayor estudio de la selección de variables en el ámbito funcional.

5.2. Trabajo futuro

Como se ha dicho, en todo el FDA hay una gran cantidad de trabajo a realizar, ya sea profundizando en las vías más asentadas como la representación de las funciones, o explorando caminos más novedosos como la aplicación de técnicas de proyecciones aleatorias [71]. En este trabajo nos hemos preocupado principalmente por la reducción de dimensiones de las funciones para su posterior clasificación. En concreto se ha probado la viabilidad y el interés de aplicar técnicas de selección de variables a datos funcionales, lo que abre varias posibles líneas de estudio a tener en cuenta.

Tanto MaxRel como mRMR se han aplicado directamente a las funciones discretizadas y los resultados son únicamente empíricos. Uno de los primeros esfuerzos de investigación debería dedi-

carse a precisar desde el punto de vista teórico el concepto de la selección de variables en el ámbito funcional y a adaptar formalmente alguno de los algoritmos de selección de variables o proponer alguno específico para funciones.

En este estudio sólo se ha experimentado con mRMR por lo que sería interesante extender las pruebas a otros métodos con distintos criterios, por ejemplo parecen adecuados la variante de mRMR propuesta en [75] o algoritmos basados en ReliefF [83]. Además, mRMR es un método de filtro y sería deseable estudiar el comportamiento de alguno de los sistemas de wrapper que obtienen buenos resultados en otros contextos, aun perdiendo independencia con el clasificador. Para mantener en parte esta independencia y mejorar los resultados se podrían estudiar aproximaciones a métodos de selección en 2 etapas [52] utilizando mRMR para una primera selección y un algoritmo de wrapper para reordenarla, ya que se ha visto en algunos casos que se obtendría un mayor acierto cambiando el orden de algunos atributos.

Al tratarse de un estudio preliminar y de una primera aproximación las pruebas no han sido exhaustivas y convendría ampliarlas. Se ha utilizado un número reducido de muestras, que aunque ilustran distintas situaciones, sería positivo completar con más conjuntos, especialmente multiclase. Los conjuntos utilizados son de alta dimensionalidad, pero no extrema, y además el número de ejemplos es equiparable o superior al de atributos, por lo que sería interesante realizar pruebas con muestras con un ratio más extremo de variables sobre el número de patrones. También sería deseable estudiar el comportamiento de mRMR variando parámetros como el tipo de discretización de los atributos, pero especialmente modificando la igualdad entre relevancia y redundancia dando más peso a una de ellas y observando los efectos. Además, se ha dejado al margen uno de los principales problemas asociados a la reducción de dimensión, que consiste en determinar la dimensión del espacio reducido. Este es un problema complejo que suele afrontarse asumiendo como bueno el momento en el que deja de producirse una mejora "significativa" al añadir una nueva variable o componente. Sin embargo, este sistema no es independiente del clasificador y otro aspecto a investigar sería la estimación de esta dimensión.

Para concluir, aunque sólo se ha tocado tangencialmente y escapa al alcance de este trabajo, otro tema interesante sería la adopción de un clasificador base, o de referencia. De los experimentos realizados no puede extraerse ninguna conclusión, ya que no estaban encaminado a tal efecto. Sin embargo se podrían tratar de comprobar o refutar las tesis propuestas por Hand [47] en cuanto a los algoritmos simples, y en concreto la utilización de KNN como algoritmo de referencia, al menos en el ámbito funcional [3].

Bibliografía

[1] ABRAHAM, C., BIAU, G., AND CADRE, B. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics 58*, 3 (2006), 619–633.

[2] ACI, M., INAN, C., AND AVCI, M. A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. *Expert Systems with Applications* (2009).

[3] BAÍLLO, A., CUESTA-ALBERTOS, J., AND CUEVAS, A. Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics (under revision)* (2010).

[4] BAILLO, A., AND CUEVAS, A. Supervised functional classification: A theoretical remark and some comparisons. *Arxiv preprint arXiv:0806.2831* (2008).

[5] BAÍLLO, A., CUEVAS, A., AND FRAIMAN, R. Classification methods for functional data. In *The Oxford Handbook of Functional Data Analysis*, F. Ferraty and Y. Romain, Eds. Oxford University Press, USA, 2011, pp. 259–297.

[6] BARKER, M., AND RAYENS, W. Partial least squares for discrimination. *Journal of Chemometrics 17*, 3 (2003), 166 – 173.

[7] BENKO, M., HÄRDLE, W., AND KNEIP, A. Common functional principal components. *Annals of Statist 37* (2009), 1–34.

[8] BERLINET, A., BIAU, G., AND ROUVIERE, L. Functional Supervised Classification with Wavelets. In *Annales de l'ISUP* (2008), vol. 52, Institut de statistique de l'Université de Paris, pp. 61–80.

[9] BERLINET, A., AND THOMAS-AGNAN, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2003.

[10] BIAU, G., BUNEA, F., AND WEGKAMP, M. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory 51*, 6 (2005), 2163–2172.

[11] BIAU, G., DEVROYE, L., AND LUGOSI, G. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory 54*, 2 (2008), 781–790.

[12] BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 ed. Springer, 2007.

[13] BLUM, A., AND LANGLEY, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence 97*, 1-2 (1997), 245–271.

[14] BOOR, C. D. *A Practical Guide to Splines*. Springer, 2001.

[15] BOULESTEIX, A.-L., AND STRIMMER, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics 8*, 1 (2007), 32–44.

[16] CARDOT, H., CRAMBES, C., KNEIP, A., AND SARDA, P. Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics & Data Analysis 51*, 10 (2007), 4832–4848.

[17] CARREIRA-PERPIÑÁN, M. A. A review of dimension reduction techniques. Tech. Rep. CS–96–09, Dept. of Computer Science, University of Sheffield, 1997.

[18] CHANG, C.-C., AND LIN, C.-J. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] CHIOU, J., AND MÜLLER, H. Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis 51*, 10 (2007), 4849–4863.

[20] COETZEE, F. M., GLOVER, E., LAWRENCE, S., AND GILES, C. L. Feature selection in web applications by roc inflections and powerset pruning. In *In Proceedings of 2001 Symp. on Applications and the Internet (SAINT 2001* (2001), IEEE Computer Society, pp. 5–14.

[21] CUESTA-ALBERTOS, J., AND FRAIMAN, R. Impartial trimmed k-means for functional data. *Computational Statistics & Data Analysis 51*, 10 (2007), 4864–4877.

[22] CUESTA-ALBERTOS, J., FRAIMAN, R., AND RANSFORD, T. A sharp form of the cramér–wold theorem. *Journal of Theoretical Probability 20* (2007), 201–209.

[23] CUEVAS, A., FEBRERO, M., AND FRAIMAN, R. Linear functional regression: The case of fixed design and functional response. *Canadian Journal of Statistics 30*, 2 (2002), 285–300.

[24] CUEVAS, A., FEBRERO, M., AND FRAIMAN, R. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics 22*, 3 (2007), 481–496.

[25] CUEVAS, A., AND FRAIMAN, R. On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis 100*, 4 (2009), 753–766.

[26] DEVROYE, L., GYÖRFI, L., AND LUGOSI, G. *A Probabilistic Theory of Pattern Recognition*, corrected ed. Springer, 1996.

[27] DIAZ-URIARTE, R., AND ALVAREZ DE ANDRES, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7* (2006), 3.

[28] DING, C., AND PENG, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology 3*, 2 (2005), 185–205.

[29] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern Classification*, 2 ed. John Wiley and sons, inc, 2003.

[30] DY, J. G., AND BRODLEY, C. E. Feature selection for unsupervised learning. *J. Mach. Learn. Res. 5* (2004), 845–889.

[31] ESCABIAS, M., AGUILERA, A., AND VALDERRAMA, M. Functional PLS logit regression model. *Computational Statistics & Data Analysis 51*, 10 (2007), 4891–4902.

[32] FEBRERO BANDE, M. A present overview on functional data analysis. *BEIO, Boletín de Estadística e Investigación Operativa 24*, 1 (2008), 6.

[33] FENG, S., KOGAN, I., AND KRIM, H. Classification of curves in 2D and 3D via affine integral signatures. *Acta applicandae mathematicae 109*, 3 (2010), 903–937.

[34] FERRATY, F., AND VIEU, P. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis 44*, 1-2 (2003), 161–173.

[35] FERRATY, F., AND VIEU, P. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[36] FERRÉ, L., AND VILLA, N. Multilayer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics 33*, 4 (2006), 807–823.

[37] FORMAN, G. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research 3* (2003), 1289–1305.

[38] FORT, G., AND LAMBERT-LACROIX, S. Classification using partial least squares with penalized logistic regression. *Bioinformatics 21* (2005), 1104–1111.

[39] FRAIMAN, R., JUSTEL, A., AND SVARC, M. Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association 103*, 483 (2008), 1294–1303.

[40] FRAIMAN, R., AND MUNIZ, G. Trimmed means for functional data. *TEST 10* (2001), 419–440.

[41] FRANK, I. E., AND FRIEDMAN, J. H. A statistical view of some chemometrics regression tools. *Technometrics 35*, 2 (1993), 109–135.

[42] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3* (2003), 1157–1182.

[43] GUYON, I., GUNN, S., NIKRAVESH, M., AND ZADEH, L. A., Eds. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, 1 ed. Springer, 2006.

[44] GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine Learning 46*, 1-3 (2002), 389–422.

[45] HALL, M. Correlation-based feature selection for machine learning, 1998.

[46] HALL, P., AND HOROWITZ, J. Bandwidth selection in semiparametric estimation of censored linear regression models. *Econometric Theory 6*, 02 (2009), 123–150.

[47] HAND, D. Classifier technology and the illusion of progress. *Statistical Science 21*, 1 (2006), 1–14.

[48] HAREZLAK, J., COULL, B., LAIRD, N., MAGARI, S., AND CHRISTIANI, D. Penalized solutions to functional regression problems. *Computational statistics & data analysis 51*, 10 (2007), 4911–4925.

[49] HASTIE, T., BUJA, A., AND TIBSHIRANI, R. Penalized discriminant analysis. *The Annals of Statistics 23*, 1 (1995), 73–102.

[50] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*, 2 ed. Springer, 2009.

[51] HLUBINKA, D., AND PRCHAL, L. Changes in atmospheric radiation from the statistical point of view. *Computational Statistics & Data Analysis 51*, 10 (2007), 4926–4941.

[52] HUA, J., TEMBE, W., AND DOUGHERTY, E. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition 42*, 3 (2009), 409–424.

[53] JAKULIN, A., AND BRATKO, I. Testing the significance of attribute interactions. In *Proc. of 21st International Conference on Machine Learning (ICML)* (Banff, Alberta, Canada, 2004), R. Greiner and D. Schuurmans, Eds., pp. 409–416.

[54] JIRAPECH-UMPAI, T., AND AITKEN, S. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics 6* (2005), 148.

[55] KAYSER, C., PETKOV, C., AUGATH, M., AND LOGOTHETIS, N. Functional imaging reveals visual modulation of specific fields in auditory cortex. *Journal of Neuroscience 27*, 8 (2007), 1824.

[56] KENT, J. T. Information gain and a general measure of correlation. *Biometrika 70*, 1 (1983), 163–173.

[57] KITTLER, J. Feature set search algorithms. In *Pattern Recognition and Signal Processing* (1978), C. Chen, Ed., Springer, pp. 41–60.

[58] KOHAVI, R., AND JOHN, G. H. Wrappers for feature subset selection. *Artif. Intell. 97*, 1-2 (1997), 273–324.

[59] KOLMOGOROV, A. N., AND FOMIN, S. V. *Elements of the Theory of Functions and Functional Analysis*. Dover Publications, 1999.

[60] LE CAO, K. A., GONCALVES, O., BESSE, P., AND GADAT, S. Selection of biologically relevant genes with a wrapper stochastic algorithm. *Stat Appl Genet Mol Biol* (2007).

[61] LI, B., AND YU, Q. Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis 52*, 10 (2008), 4790–4800.

[62] LI, L. Dimension reduction for high-dimensional data. *Methods Mol. Biol. 620* (2010), 417–434.

[63] LIU, H., AND MOTODA, H., Eds. *Computational Methods of Feature Selection (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman and Hall/CRC, 2007.

[64] LIU, H., AND YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering 17*, 4 (2005), 491–502.

[65] LIU, Y., AND RAYENS, W. Pls and dimension reduction for classification. *Computational Statistics 22*, 2 (2007), 189–208.

[66] LÓPEZ-PINTADO, S., AND ROMO, J. On the concept of depth for functional data. *Journal of the American Statistical Association 104*, 486 (2009), 718–734.

[67] MALDONADO, S., AND WEBER, R. A wrapper method for feature selection using support vector machines. *Information Sciences 179*, 13 (2009), 2208–2217.

[68] MANTEIGA, W. G., AND VIEU, P. Statistics for functional data. *Computational Statistics & Data Analysis 51*, 10 (2007), 4788–4792.

[69] NERINI, D., AND GHATTAS, B. Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis 51*, 10 (2007), 4984–4993.

[70] NGUYEN, D. V., AND ROCKE, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics 18* (2002), 39–50.

[71] NIETO REYES, A. *Aplicaciones estadísticas de las proyecciones aleatorias*. PhD thesis, Universidad de Cantabria, 2010.

[72] PENG, H., LONG, F., AND DING, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell 27* (2005), 1226–1238.

[73] PHATAK, A., AND DEHOOG, F. Exploiting the connection between plsr, lanczos, and conjugate gradients: alternative proofs of some properties of plsr. *Journal of Chemometrics 16*, 3 (2002), 361 – 7.

[74] POLLARD, D. Strong consistency of k-means clustering. *The Annals of Statistics 9*, 1 (1981), 135–140.

[75] PONSA, D., AND LÓPEZ, A. Feature selection based on a new formulation of the minimal-redundancy-maximal-relevance criterion. In *Pattern Recognition and Image Analysis*, vol. 4477 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007, pp. 47–54.

[76] PREDA, C. Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of statistical planning and inference 137*, 3 (2007), 829–840.

[77] PREDA, C., SAPORTA, G., AND LÉVÉDER, C. PLS classification of functional data. *Comput. Stat. 22*, 2 (2007), 223–235.

[78] PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., AND FLANNERY, B. P. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3 ed. Cambridge University Press, 2007.

[79] RAMSAY, J., HOOKER, G., AND GRAVES, S. *Functional Data Analysis with R and MATLAB (Use R)*. Springer, 2009.

[80] RAMSAY, J., AND SILVERMAN, B. *Applied Functional Data Analysis*, 1 ed. Springer, 2002.

[81] RAMSAY, J., AND SILVERMAN, B. W. *Functional Data Analysis (Springer Series in Statistics)*, 2nd ed. Springer, 2005.

[82] REFAEILZADH, P., TANG, L., AND LUI, H. On comparison of feature selection algorithms. In *AAAI 2007 Workshop on Evaluation Methods for Machine Learning II* (2007), pp. 1–6.

[83] ROBNIK-ŠIKONJA, M., AND KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning 53*, 1-2 (2003), 23–69.

[84] RODRIGUEZ-LUJAN, I., HUERTA, R., ELKAN, C., AND CRUZ, C. S. Quadratic programming feature selection. *Journal of Machine Learning Research 11* (2010), 1491–1516.

[85] ROSIPAL, R., AND TREJO, L. J. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research 2* (2001), 97–123.

[86] ROSSI, F., AND CONAN-GUEZ, B. Theoretical properties of projection based multilayer perceptrons with functional inputs. *Neural Processing Letters 23*, 1 (2006), 55–70.

[87] ROSSI, F., AND VILLA, N. Support vector machine for functional data classification. *Neurocomputing 69*, 7-9 (2006), 730–742.

[88] SAEYS, W., DE KETELAERE, B., AND DARIUS, P. Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics 22*, 5 (2008), 335–344.

[89] SAEYS, Y., INZA, I., AND LARRANAGA, P. A review of feature selection techniques in bioinformatics. *Bioinformatics 23* (2007), 2507–2517.

[90] SHIMIZU, N., AND MIZUTA, M. Functional clustering and functional principal points. In *Knowledge-Based Intelligent Information and Engineering Systems* (2010), Springer, pp. 501–508.

[91] SINGHI, S. K., AND LIU, H. Feature subset selection bias for classification learning. In *ICML '06: Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 849–856.

[92] TIAN, T., AND JAMES, G. Interpretable Dimensionality Reduction for Classification with Functional Data. *Under revision* (2010).

[93] TRYGG, J. *Parsimonious multivariate models*. PhD thesis, Umetrics Academy, Umea, 2001.

[94] VAN DER MAATEN, L. J. P., POSTMA, E. O., AND VAN DEN HERIK, H. J. Dimensionality reduction: A comparative review. 2007.

[95] VAPNIK, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[96] VARDI, Y., AND ZHANG, C. The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America 97*, 4 (2000), 1423.

[97] VARSHAVSKY, R., GOTTLIEB, A., LINIAL, M., AND HORN, D. Novel unsupervised feature filtering of biological data. *Bioinformatics 22* (2006), e507–513.

[98] VEGA VILCA, J. C. *Generalizaciones de mínimos cuadrados parciales con aplicación en clasificación supervisada*. PhD thesis, Universidad de Puerto Rico, 2004.

[99] VILLA, N., ROSSI, F., AND CARCASSONNE, F. Recent advances in the use of SVM for functional data classification. In *Functional and Operatorial Statistics: Proceedings of 1rst International Workshop on Functional and Operatorial Statistics (IWFOS 2008), Contributions to Statistics*, pp. 273–280.

[100] WANG, S., JANK, W., AND SHMUELI, G. Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics 26*, 2 (2008), 144–160.

[101] WEI, L., AND KEOGH, E. Semi-supervised time series classification. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), ACM, pp. 748–753.

[102] WENG, X., AND SHEN, J. Time Series Classification Using Locality Preserving Projections. In *2007 IEEE International Conference on Automation and Logistics* (2007), pp. 1392–1397.

[103] WESTON, J., MUKHERJEE, S., CHAPELLE, O., PONTIL, M., POGGIO, T., AND VAPNIK, V. Feature selection for svms. In *NIPS* (2000), T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., MIT Press, pp. 668–674.

[104] XIONG, M., FANG, X., AND ZHAO, J. Biomarker identification by feature wrappers. *Genome Res. 11* (2001), 1878–1887.

[105] YAMPOLSKIY, R., AND GOVINDARAJU, V. Behavioural biometrics: a survey and classification. *International Journal of Biometrics 1*, 1 (2008), 81–113.

[106] YAO, F., MÜLLER, H., AND WANG, J. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*, 470 (2005), 577–590.

[107] YOUNES, Z., ABDALLAH, F., AND DENOEUX, T. Multi-label classification algorithm derived from k-nearest neighbor rule with label dependencies. In *roceedings of the 16th European Signal Processing Conference* (2008).

[108] YU, B., OSTLAND, M., GONG, P., AND PU, R. Penalized discriminant analysis of in situ hyperspectral data forconifer species recognition. *Ieee transactions on Geoscience and remote sensing 37*, 5 Part 2 (1999), 2569–2577.

[109] YU, L., AND LIU, H. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res. 5* (2004), 1205–1224.

[110] ZENG, X.-Q., WANG, M.-W., AND NIE, J.-Y. Text classification based on partial least square analysis. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing* (New York, NY, USA, 2007), ACM, pp. 834–838.

[111] ZHANG, M., AND ZHOU, Z. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition 40*, 7 (2007), 2038–2048.

[112] ZHAO, Z., AND LIU, H. Searching for interacting features in subset selection. *Intell. Data Anal. 13*, 2 (2009), 207–228.

[113] ZUO, Y., AND SERFLING, R. General notions of statistical depth function. *Annals of Statistics 28*, 2 (2000), 461–482.